ChinaRxiv 中X

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-202307.00301

Complex Network Construction and Intelligent Application Exploration in University Libraries (Postprint)

Authors: Shi Guoliang, Xie Zeyu, Yang Xiaoli

Date: 2023-07-26T00:00:00+00:00

Abstract

[目的/意义] University libraries exhibit high levels of informatization, yet their capabilities in data mining and intelligentization remain to be enhanced. Complex networks utilize graph databases as the carrier for storage and graph queries, enabling unified organization and mining of graph-structured data. Graph embedding and graph algorithm technologies, compared to traditional machine learning methods, can fully mine the implicit relationships within graph-structured data. This study employs complex network technology to integrate multi-source data and investigates the role of graph-structured data mining methods, including graph embedding technology and graph algorithms, in enhancing the intelligentization level of libraries. [方法/过程] First, data feature analysis and cleaning are performed based on accessible data. Second, a complex network conceptual model is constructed according to data features, and network construction and storage are implemented using Neo4j batch import technology. Finally, the application of graph algorithms and graph embedding technologies in graph-structured data mining is explored. [结果/结论] A library complex network is constructed by integrating multi-source data using a graph structure, with a graph database serving as the storage medium. Graph algorithms and graph embedding technologies possess unique advantages in library intelligent applications such as user profiling analysis, precision recommendation, and intelligent question-answering.

Full Text

Preamble

Vol. 63 No. 23, December 2019
Exploring the Construction and Intelligent Applications of Complex Networks in University Libraries

ChinaRxiv 中X

Shi Guoliang[1], Xie Zeyu[1], Yang Xiaoli[2]
[1] Business School, Hohai University, Nanjing 211100
[2] Hohai University Library, Nanjing 211100

Abstract

[**Purpose/Significance**] University libraries exhibit high levels of informatization, yet their data mining and intelligence capabilities require further enhancement. Complex networks utilize graph databases as carriers for storage and graph querying, enabling unified organization and mining of graph-structured data. Graph embedding and graph algorithm techniques can fully exploit implicit connections within graph-structured data, offering advantages over traditional machine learning methods. This study employs complex network technology to integrate multi-source data and explores the role of graph-structured data mining methods—such as graph embedding and graph algorithms—in advancing library intelligence.

[**Method/Process**] First, we analyze and clean available data based on its characteristics. Second, we construct a complex network conceptual model tailored to the data features and implement network construction and storage using Neo4j batch import technology. Finally, we explore applications of graph algorithms and graph embedding techniques in graph-structured data mining.

[**Result/Conclusion**] Multi-source data is integrated using graph structures to construct a library complex network, with graph databases serving as the storage medium. Graph algorithms and graph embedding technologies offer unique advantages in intelligent library applications such as user profiling analysis, precision recommendation, and intelligent question answering.

Classification Number: G254
Keywords: complex network, graph database, graph algorithms, graph embedding, intelligent library
DOI: 10.13266/j.issn.0252-3116.2019.23.012

Introduction

The 13th Five-Year Plan (2016) and the 2017 Government Work Report both emphasized the need to vigorously develop artificial intelligence [1]. In recent years, research on machine learning, deep learning, and knowledge graphs has intensified, with enterprises rapidly following suit. The rapid development of AI has been driven by improvements in data collection and computing power, as well as advances in data mining methods such as deep learning and machine learning. Against this backdrop of big data and AI, library services are transitioning from informatization to intelligence [2]. The comprehensive application of AI technologies to explore intelligent applications in library services can help better meet reader needs and improve the efficiency of library resource utilization and service effectiveness.

Libraries have begun using industrial robots for book reception, classification, and registration [4]. Self-service libraries have been deployed across China. For instance, the Capital Library has developed a reference consultation robot, while Shanghai Jiao Tong University Library's "Xiao Jiao" can provide information services and engage in autonomous conversations with readers [5]. In Dongguan, RFID-enabled mini self-service libraries have been implemented to provide citizens with 24-hour book borrowing and return services [5]. Although the level of intelligent application in libraries still needs improvement, these applications have significantly promoted the efficiency of book resource utilization and library service levels.

Currently, the level of big data analysis and intelligent application in libraries struggles to meet user demands for precision recommendation and knowledge services [3]. This study combines library big data, uses graph database technology to construct complex networks, and explores the role of graph algorithms and graph embedding technologies in intelligent applications such as user profiling, recommendation systems, and intelligent question answering. The integration of complex networks, graph databases, graph algorithms, and graph embedding technologies into library data analysis and mining can help libraries better serve readers and improve resource utilization efficiency.

1 Research Status

1.1 Library Intelligent Services

Since the 1970s, intelligent technologies have gradually been applied to library management and services. In 1994, the municipal library of Örnsköldsvik, Sweden, pioneered the use of intelligent library systems. Despite high levels of informatization, libraries face significant challenges in transitioning from informatization to intelligence [6]. On one hand, library intelligent services require specialized talent and substantial financial support. Professionals skilled in technical methods are needed to improve library service quality, and the research exploration of transforming libraries from informatization to intelligence cannot proceed without funding. On the other hand, libraries have deficiencies in current data collection and service management. Data is the cornerstone of intelligent applications, and libraries require extensive software and hardware infrastructure to collect and manage big data. Data collection and usage must comply with regulations, and user privacy must be protected without infringing upon user rights. As centers for knowledge, learning, and communication, libraries can achieve tremendous value by employing emerging data mining technologies and AI methods to enhance service levels and resource utilization efficiency.

1.2 Complex Networks and Graph Databases

Networks are ubiquitous in nature and human society, including social networks, the Internet, power grids, and airline networks. Complex network the-

ory explains these widespread network phenomena and their complexity, typically exhibiting small-world properties, hierarchical characteristics, and self-organization [7]. Wu Zhiqin et al. conducted user profiling research for university libraries based on social network analysis [8]. Feng Lei et al. proposed incorporating complex network theory into library personalized recommendation services [9]. Zhao Peng et al. applied complex network features for document keyword extraction in natural language processing [10]. Zhai Dongsheng et al. utilized graph databases in patent knowledge base construction [11]. Li Hui et al. studied trust propagation analysis in complex network environments [12]. Complex network theory and research methods have broad applications in library digital resource integration, citation networks, and scientific collaboration networks [13]. Li Deyi et al. observed that complex network theory has received increasing attention in research exploring human cognition and thinking mechanisms, with attempts to construct cognitive models with complex network characteristics to represent uncertain cognitive processes [14].

Complex networks rely on graph databases for storage and application, with both data storage structures and query methods based on graph theory. Neo4j, the most widely used enterprise-level graph database, implements a property graph model that can describe most graph usage scenarios using nodes and relationships as fundamental data structures. Graph database-based data analysis and mining have mature applications in many fields, such as anti-fraud model research in banking and insurance, social relationship network construction for relationship recommendations at LinkedIn, and logistics network structure optimization using graph algorithms to reduce transportation costs [15]. As graph database technology advances, it not only enables the storage of associated networks but also integrates numerous classical graph algorithms such as community detection, centrality algorithms, and path planning. Graph algorithms facilitate in-depth analysis and research of network structures. Graph embedding technology maps nodes in network structures to low-dimensional dense vectors, further expanding the means of graph-structured data mining [16]. The comprehensive use of complex networks and graph databases to integrate multi-source data enables targeted analysis and mining of associated data, thereby enhancing library data mining levels and improving intelligent service capabilities.

2 Research Design

Complex networks aim to integrate library borrowing data into interconnected data networks. Data mining based on associated data networks demonstrates significant advantages over table storage structures, substantially improving intelligent application levels through network structure analysis. The process of constructing library complex networks involves: (1) determining data boundaries and requirements; (2) analyzing features of various extracted data and establishing data cleaning rules based on the complex network conceptual model; (3) cleaning data according to established rules while protecting reader privacy; (4) applying Neo4j batch import technology for node and relationship construc-

tion; (5) conducting log checks, indexing, and query optimization on the imported graph database; and (6) exploring intelligent applications on the constructed and optimized complex network. This process is illustrated in Figure 1 [Figure 1: see original paper].

Complex networks consist of numerous interconnected nodes and relationships. The underlying conceptual model design determines the network structure, which should comprehensively consider application scenarios and graph database performance. This study constructs a complex network integrating book information data, reader information data, and reader borrowing data. The conceptual model design follows two principles: first, it must represent the entire relationship network; second, it must optimize subgraph query efficiency in the graph database. Neo4j's underlying storage features include: (1) nodes and relationships; (2) properties for both nodes and relationships; (3) nodes with one or more labels, while relationships have only one type; and (4) directed relationships from one node to another [17]. The network structure can be abstracted as a network composed of entities such as readers, books, and authors connected through borrowing and writing relationships. The complex network conceptual model is shown in Figure 2 [Figure 2: see original paper].

The hierarchical structure of nodes and relationships in complex networks is illustrated in Figure 3 [Figure 3: see original paper]. The entire complex network comprises nodes (readers, books, authors) and relationships (borrowing, writing). Each node has properties, primary keys, and labels. For example, reader nodes store attributes such as student ID, name, gender, college, total books borrowed, credit status, and reader type, with the student ID serving as the unique primary key identifier. Nodes can have multiple labels, and each relationship can also store properties and types. This transformation from table structure to graph structure data enables full exploitation of graph data mining advantages in association analysis.

3 Complex Network Construction Process

3.1 Data Sources

The data source for this study is the book borrowing system of Library A, originally stored in an Oracle relational database. The data tables were designed according to existing system workflows, resulting in data separation, redundancy, and diverse types. This study uses graph database technology to transform table-structured data into associated network-structured data and conducts analysis and mining based on graph algorithms and graph embedding techniques. The data primarily consists of three components: reader information (name, major, grade, etc.), book information (title, classification number, author, publication year, etc.), and book borrowing data (time, browsing records, borrowing records, etc.). The extracted data spans from 2008 to 2018, involving 34 data tables.

3.2 Data Cleaning

The design of the borrowing network conceptual model fully considers the representation capability of the network and query efficiency in practical use (see Figure 2). Since Neo4j's underlying storage structure consists of nodes and relationships, the transformation from relational database table storage to graph storage is necessary. Relational databases involve multi-table operations and deep queries when describing complex associations, resulting in far lower efficiency than graph queries.

3.3 Complex Network Construction and Storage

The complex network is constructed by transforming the cleaned relational database data into graph structures according to the conceptual model. Multiple methods exist for importing data into graph databases, including Cypher CREATE statements, Cypher LOAD CSV statements, Java API, neo4j-import, and neo4j-apoc-load. Using the neo4j-import tool to import prepared CSV files achieves speeds of 120,000 (nodes + relationships) per second with minimal resource consumption. The batch import command is shown in Figure 4 [Figure 4: see original paper].

The batch import process took 19 seconds and 949 milliseconds, successfully importing 715,682 nodes, 1,146,925 relationships, and 2,861,027 properties. After completion, a Lib0818.db graph database file is generated in the specified path. Neo4j integrates a visualization frontend. By starting the Neo4j service, modifying the configuration file to specify the database location, and entering the URL (http://localhost:7474/browser/) in a browser, users can connect to the designated database for querying and visualization operations. The entire complex network consists of nodes and edges forming a tightly connected relationship network. Based on the conceptual model described in Figure 2, we extract table-structured data from the Oracle database, conduct feature analysis, clean and organize CSV files, use neo4j-import for batch data import, and implement query and visualization operations through the frontend page, as shown in Figure 5 [Figure 5: see original paper].

3.4 Graph Database Optimization

To maximize graph database advantages, data must be more finely partitioned and deeply labeled. Integrating multi-source data through graph structures can uncover more information value, with larger datasets revealing more subtle associations in complex networks. As the carrier for complex network storage and application, graph databases require optimal query efficiency. Cypher is Neo4j's query language, offering both readability and optimal performance. Five optimization measures are recommended [15]: (1) deep labeling to avoid global data scanning; (2) introducing indexing and constraint mechanisms (Cypher statements for complex network indexes and constraints are shown in Figure 6 [Figure 6: see original paper]); (3) avoiding Cartesian product operations and

unexpected query results; (4) query performance analysis using the PROFILE keyword to monitor detailed query processing information; and (5) graph model optimization with clear model definitions for rapid query matching.

4 Complex Network Application Exploration

Complex network technology integrates multi-source heterogeneous data from different sources and standards for unified management, enabling data mining to achieve a "1+1>2" effect. Graph-structured data analysis and mining primarily employ two approaches: graph algorithms (e.g., community detection, PageRank) and graph embedding techniques that map network nodes to low-dimensional dense vectors, enabling machine learning and deep learning methods beyond traditional graph algorithms. In library big data, complex networks can fuse reader personal information, browsing records, and borrowing records for comprehensive analysis, enhancing system understanding of users, accurately grasping reader attributes and preferences, and enabling applications in user profiling, personalized recommendation, and knowledge-based question answering.

4.1 User Profiling

User profiling is a collection of user information images obtained from massive data [19]. User profiling research helps libraries accurately understand reader needs and holds significant value in improving library service quality and precision marketing [20]. Libraries possess extensive data resources on users, books, and interactions, enabling the use of data mining technologies to organize, analyze, and mine data, and construct reader virtual profiles based on integrated multi-source data [21]. Associated data encompassing more comprehensive reader information facilitates precise extraction of user tags and more accurate reader characterization.

In library contexts, user profiling faces several challenges: data sparsity and separation, lack of interconnectivity with external data, and data mining methods that analyze single data sources without fusing multi-source data to leverage associations. Complex networks connect these "data islands," integrating multi-source data for unified organization and analysis to discover associations that single data sources cannot reveal. Figure 7 [Figure 7: see original paper] illustrates the extraction approach for reader user profiling under library data fusion, where multi-source data includes not only reader personal attributes and book attributes but also borrowing records and browsing query records generated from each interaction. Network structures are analyzed and mined using graph algorithms—community detection algorithms enable user clustering, while centrality algorithms (e.g., PageRank) enable analysis of reader node influence. In addition to Neo4j's integrated graph algorithms, open-source libraries such as APOC can assist in algorithm implementation. Using graph algorithms for network structure data mining under data association enables more accurate identification of latent associations and characteristics among readers.

4.2 Personalized Recommendation

Traditional book recommendation systems primarily rely on collaborative filtering and content-based recommendation. Collaborative filtering recommends items based on user rating similarity—users who like similar items share similar preferences—yet most university libraries lack rating data, possessing only borrowing records. Content-based recommendation struggles to achieve dynamic, personalized recommendation lists due to the relatively stable attributes of readers (major, college) and books (category, classification number) [22]. Personalized recommendation has distinct characteristics in different domains; university library book recommendation features include: books as recommendation objects with broad coverage and predominantly professional titles, and a stable service population of faculty and students with high specialization and easy clustering [23]. Current university library recommendation systems suffer from low personalization and insufficient data collection and mining.

Complex network-based book recommendation strategies include: (1) recommendation based on subgraph queries and custom rules, and (2) entity vectorization through graph embedding for recommendation based on entity similarity. The rule-based approach offers advantages including real-time dynamic recommendation, cold start mitigation, flexible rules, good result interpretability, and enhanced user experience. Figure 8 [Figure 8: see original paper] illustrates rule-based complex network recommendation strategies, including similarity recommendation based on the same author, content-based recommendation through book classification analysis, and clustering recommendation by discovering reading lists of readers with similar preferences. However, manually defined rules struggle to adapt to all scenarios and exhibit high computational complexity in large-scale complex networks, making efficiency difficult to guarantee. Graph embedding technology addresses this by mapping network nodes to low-dimensional dense vectors for recommendation. Many technology companies currently employ graph embedding for large-scale recommendation [24], with mature algorithms including DeepWalk [24], node2vec [25], SDNE [26], and LINE [27]. Using graph embedding technology to vectorize nodes enables various applications through node similarity calculation, with personalized recommendation being the most typical.

4.3 Intelligent Question Answering

Against the backdrop of advancing AI technologies, question answering systems form the technical foundation for intelligent library reference consultation services [28]. Traditional retrieval uses string matching and ranking algorithms to display relevant information pages, whereas intelligent question answering provides direct accurate answers based on question understanding. Knowledge graphs serve as knowledge carriers and the foundation for question answering systems. In addition to general knowledge bases like DBpedia, YAGO, and Freebase, domain knowledge bases feature fine granularity, high quality, and complex reasoning. Constructing domain knowledge bases from library big data

and storing knowledge using graph database technology can support intelligent consultation.

Current question answering system implementations include two primary approaches: (1) graph matching-based methods that represent semantic relationships in natural language as graphs, transforming natural language questions into subgraph matching problems [29]. Figure 9 [Figure 9: see original paper] illustrates the knowledge question answering flow based on graph matching using complex networks as the knowledge base carrier: first, process natural language questions to extract entities; second, match corresponding question templates based on entity types and transform the original question into subgraph query statements; finally, query the knowledge base using the generated statements and return results to users. Due to diverse natural language grammar and complex expressions, accurate entity detection via subgraph matching requires extensive labeled data and manually defined rules, hindering scalability. (2) Vectorized representation-based question-answer pair matching methods. Vectorized node representation applies not only to recommendation but also to knowledge question answering. QA datasets are prepared to train vectorized representations of nodes and relationships, mapping answers and questions into the same vector space. Question-answer matching then requires only similarity calculation between vectorized representations, without considering question syntax and semantics. This approach has achieved significant progress in general knowledge graph intelligent question answering research [30-31]. Another implementation method is vectorized representation of graph structures, including node vectorization, path vectorization, and subgraph vectorization, using vector similarity calculation for knowledge question answering. Vectorized representations containing more information enable more accurate answers.

References

[1] Wu Jianzhong. Artificial Intelligence and Libraries[J]. Library and Information, 2017(6): 1-5.
[2] Li Caining, Bi Xinhua, Chen Lijun. Research on Smart Library Service Models and Platform Construction[J]. Library, 2018(12): 1-7.
[3] Buhe Baolide. Application, Challenges, and Development Trends of Artificial Intelligence Technology in Libraries[J]. Library and Information, 2017(6): 48-54.
[4] Liu Xiaomin. Library—A New Era of Automation[J]. Robot Technology and Application, 1997(1): 7-8.
[5] Wang Zhanni, Zhang Guoliang. A Review of Library Robot Applications[J]. Journal of Academic Libraries, 2015, 33(3): 82-87.
[6] Yang Jiulong, Yang Yuxuan, Xu Bi han. Theoretical Logic, Realistic Dilemmas, and Path Prospects of Artificial Intelligence in Libraries[J]. Library and Information Service, 2019, 63(4): 32-39.
[7] Wang Xiaofan, Li Xiang. Complex Network Theory and Its Applications[M]. Beijing: Tsinghua University Press, 2006: 18-46.

[8] Wu Zhiqin, Liu Yijun, Li Renpu, et al. Research on User Profile Construction for University Libraries Based on Social Networks[J]. Library Science Research, 2018(16): 26-30.
[9] Feng Lei, Zhang Yuguang, Tang Li. Application of Complex Network Theory in Library Personalized Recommendation Services[J]. Information Studies: Theory & Application, 2009, 32(2): 69-71.
[10] Zhao Peng, Cai Qingsheng, Wang Qingyi, et al. A Chinese Document Keyword Extraction Algorithm Based on Complex Network Features[J]. Pattern Recognition and Artificial Intelligence, 2007, 20(6): 827-831.
[11] Zhai Dongsheng, Liu He, Zhang Jie, et al. Research on Patent Semantic Knowledge Base Construction Technology Based on Graph Databases[J]. New Technology of Library and Information Service, 2016(12): 66-75.
[12] Li Hui, Ma Xiaoping, Shi Nian, et al. Research on Trust Propagation-Based Recommendation Model in Complex Network Environment[J]. Acta Automatica Sinica, 2018, 44(2): 363-376.
[13] Li Xiaoying. Research on Complex Network Theory and Its Application in Library and Information Science[J]. Information Science, 2016, 34(10): 95-98.
[14] Li Deyi, Liu Changyu, Du Yi, et al. Uncertainty Artificial Intelligence[J]. Journal of Software, 2004(11): 1583-1594.
[15] Zhang Zhi, Pang Guoming, Hu Jiahui, et al. The Definitive Guide to Neo4j[M]. Beijing: Tsinghua University Press, 2017: 22-38.
[16] Liu Zhiyuan, Sun Maosong, Lin Yankai, et al. Research Progress on Knowledge Representation Learning[J]. Journal of Computer Research and Development, 2016, 53(2): 247-261.
[17] Kemper C. Managing Your Data in Neo4j[M]. Berkeley, CA: Apress, 2015: 57-67.
[18] Liu Haiou, Ya Sumei, Huang Wenna, et al. Contextualized Recommendation of Library Big Data Knowledge Services Based on User Profiles[J]. Library Science Research, 2018(24): 57-63.
[19] Liu Haiou, Sun Jingjing, Chen Jing, et al. User Profile Models and Their Application in Libraries[J]. Library Theory and Practice, 2018(10): 92-97.
[20] Chen Tianyuan. Empirical Study on User Profile Construction for University Mobile Libraries[J]. Library and Information Service, 2018, 62(7): 38-46.
[21] Lyu Danyang. Research on Personalized Book Recommendation Method for University Libraries Based on Association Graphs[D]. Wuhan: Huazhong University of Science and Technology, 2016.
[22] Li Min, Wang Yingchun, Liu Yanquan. Investigation and Analysis of Collection Resource Recommendation Systems in "211 Project" University Libraries[J]. Library and Information Service, 2016, 60(9): 55-60.
[23] Wang J, Huang P, Zhao H, et al. Billion-scale Commodity Embedding for E-commerce Recommendation in Alibaba[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2018: 839-848.
[24] Grover A, Leskovec J. node2vec: Scalable Feature Learning for Networks[C]//Proceedings of the 22nd ACM SIGKDD International Conference

on Knowledge Discovery and Data Mining. New York: ACM, 2016: 855-864.
[25] Wang D, Cui P, Zhu W. Structural Deep Network Embedding[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016: 1225-1234.
[26] Tang J, Qu M, Wang M, et al. LINE: Large-scale Information Network Embedding[C]//Proceedings of the 24th International Conference on World Wide Web. New York: ACM, 2015: 1067-1077.
[27] Lai Yun. Research on Library Intelligent Consultation Robot System Design and Corpus Technology[J]. Modern Information, 2017, 37(11): 121-124.
[28] Shen Kuilin, Shao Bo, Zhao Hua. Building Library Intelligent Question Answering System Using WeChat[J]. Library Science Research, 2015(8): 75-80.
[29] Bordes A, Weston J, Usunier N. Open Question Answering with Weakly Supervised Embedding Models[C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Berlin, Heidelberg: Springer, 2014: 165-180.
[30] Bordes A, Chopra S, Weston J. Question Answering with Subgraph Embeddings[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha: EMNLP, 2014: 615-620.

Author Contributions

Shi Guoliang: Proposed research proposition, defined research framework, revised manuscript.
Xie Zeyu: Conducted experiments according to research framework, wrote manuscript.
Yang Xiaoli: Guided experimental process, refined research framework, revised manuscript.

English Abstract

[Purpose/significance] The informatization level of university libraries is high, but the level of data mining and intelligence needs to be improved. The complex network uses graph database as the carrier of storage and graph query to organize and mine graph structure data. Compared with traditional machine learning methods, graph embedding and graph algorithm techniques can discover hidden connections in graph structure data. This study uses complex network to integrate multi-source data and explores the role of graph data mining methods such as graph embedding and graph algorithms in improving library intelligence. **[Method/process]** First of all, this study clarifies and analyzes the characteristics of the database based on the available data. Secondly, combined with the characteristics of data, construct a complex network conceptual model, and use Neo4j batch import technology to realize network construction and storage. Finally, explore the application of graph algorithm and graph embedding technology in graph structure data mining. **[Result/conclusion]** The multi-source data is combined with the graph structure to construct the complex network of the library, and the graph database is

used as the storage medium. Graph algorithm and graph embedding technology have unique advantages in user image analysis, accurate recommendation, intelligent QA, and other intelligent applications of the library.

Keywords: complex network; graph database; graph algorithms; graph embedding; intelligent library

Note: Figure translations are in progress. See original paper for figures.

Source: ChinaXiv — Machine translation. Verify with original.