Research on the Application of Knowledge Graph in Academic Resource Discovery Service

Jianfeng Zhou [1], Weiming Yang [2], Jing Ma [3]

[1]Library, Guangdong University of Foreign Studies, Guangzhou 510420, China. Email: 201110082@oamail.gdufs.edu.cn.

[2]Library, Guangdong University of Foreign Studies, Guangzhou 510420, China. Email: 201710076@oamail.gdufs.edu.cn

[3]Guangdong Institute of International Strategy, Guangdong University of Foreign Studies, Guangzhou 510420, China. Email: ma0703@126.com. Corresponding author.

Abstract. [Objective] To apply user graph to the service process of academic resource discovery, alleviate the dilemma of knowledge loss, and provide reference for the construction and application of user graph in university libraries. [Process] Taking academic resource discovery service as the application scenario, this paper analyzes the source, structure and characteristics of university library user data, and defines the user graph construction process of "knowledge extraction - knowledge fusion - knowledge snapshot - graph tailoring". The project of embedding user graph into academic resource discovery service process was discussed in three stages: before, during and after, and user behavior data was used for analysis and testing. [Results] The experiment proved that user graph can optimize the input and output of academic resource discovery service, reduce the cost of user behavior, and provide a reference for the application of knowledge graph.

Keywords: University library; knowledge map; academic resources; discovery; individuation

1 Introduction

Resource retrieval is a core service of university libraries and one of the main ways for users to access academic resources. With the development of diverse, shared, and massive academic resources, resource retrieval has evolved from traditional "bibliographic retrieval platforms" and "consortium-based retrieval platforms" to "academic resource discovery platforms" that provide one-stop access to resources[1]. Existing academic resource discovery platforms mainly focus on integrating heterogeneous academic resources, and they output academic resources that meet retrieval criteria through global text matching. The large number of search results often leads users to information overload and knowledge confusion. How to incorporate user characteristics into the discovery service process and meet personalized and accurate user needs has become a hot topic in this field.

Knowledge Graph (KG), as one of the mainstream tools for information organization, management and understanding in the network era, greatly improves users' experience in information diversity and result accuracy with its open organization capability and powerful semantic processing capability [2]. On the basis of knowledge graph technology, this paper

FFIT 2023, July 07-09, Chongqing, People's Republic of China

DOI 10.4108/eai.7-7-2023.2338061

proposes the user graph construction process of "knowledge extraction - knowledge fusion - knowledge snapshot - graph tailoring" for university library users. The user graph is embedded in the academic resource discovery service scenario, and its effectiveness is proved by experiments.

2 Construction of university library user map

The focus of academic resource discovery service optimization lies in mining the relationship between users and academic resources and even academic knowledge, which is implied in the user's behavior. User behaviors are stored in various service systems of libraries. These structured or semi-structured data contain information about users' personal characteristics, social relationships, knowledge needs, etc., and are the basis for libraries to perceive users. Based on the analysis of university user data, this paper proposes a user-centered user graph construction process of "knowledge extraction - knowledge fusion - knowledge snapshot - graph tailoring".

2.1 Knowledge extraction

Taking the library where the author works as an example, the main sources of user data include the user information management system, the book integrated management system, and the digital resources off-campus access system, which store the basic information of users, borrowing records, and off-campus digital resources access records respectively. The data types, contents and collection methods are shown in Table 1.

Table 1 Data types, contents and collection methods of university library users

Data type		Contents	Acquisition mode
Basic information	Social information	Age, region of origin, gender	Database import
	School information	Campus card number, status type, grade, college, major, class	
Behavioral information	Reading behavior	Person code, loan time, return time, book code	Database import
	Resource access behavior	Personnel code, data source, behavior, occurrence time, resource links	
Resource information	Books	Title, author, publisher, publication year, book series, subject words, introduction, Chinese photo classification number	Database import, Semantic model processing
	Digital resource	Title, author, author unit, keywords, subject words, publication unit, publication time, etc	Data crawler, Semantic model processing

2.1.1 User basic information knowledge extraction

The basic information of users is divided into social information and campus information. Social information is given by civil affairs units and has a wide range of social recognition, while campus information is given by colleges and universities, marking the personal characteristics of users within the campus. The basic information of users is generally stored in a structured form and can

be directly imported from the database. In order to protect the privacy of users, it is necessary to delete or partially hide the information directly related to personal privacy, such as ID number, name, mobile phone, home address, etc.

2.1.2 User behavior information knowledge extraction

User behavior mainly includes borrowing books, browsing and downloading digital resources. These behavior records are stored in a structured database in the form of "user code-behavior-academic resource code or URL" triplet, but the behavior records generally do not contain the title and full text information of academic resources, and cannot completely present the academic knowledge network. Therefore, tools such as ETL and data crawler are still needed to obtain specific information of academic resources from integrated management systems or third-party platforms.

For academic resources, in addition to the title information of authors, units, publishers, etc., unstructured information such as title, abstract, full text, pictures, audio and video also contains semantic knowledge describing the content, which needs to be extracted by a variety of semantic models. At the same time, third-party knowledge bases should be used to enrich the attributes and relationship networks of entities. In order to fully reveal the entity and improve the efficiency of knowledge fusion.

2.2 Knowledge fusion

Knowledge fusion refers to the alignment of entities with different expressions but the same meaning, such as the author's English name and Chinese name, iphone and iPhone. According to entity type, it can be divided into two parts: user knowledge fusion and academic knowledge fusion.

User entities are mainly integrated with campus card number and ID number as unique codes, and their social information and campus information entities have corresponding management standard documents, such as gender division into male and female, domiciles in accordance with the division of national administrative regions, and majors in accordance with the Catalog of Undergraduate Majors in Ordinary colleges and Universities, etc. Therefore, management standard documents can be used as prior knowledge to establish a rule dictionary. Alignment fusion based on text consistency.

For academic resource entities, the system code or URL is generally used as a unique identifier for integration, such as book ISBN, download URL of digital resources, etc. For academic knowledge entities with different concepts such as subject words, authors and publishers, resource entities are aligned based on the consistency of entity attributes or associated entity combinations [3]. Secondly, for the factual entities such as the publishing unit and the author unit, the prior knowledge is used to build a data dictionary to realize the alignment and fusion of the publishing unit and the author unit. For author entities with diversified expression forms, third-party knowledge base [4] can be introduced to confirm them. Finally, for semantic knowledge such as keywords and subject headings, on the one hand, the co-occurrence of entities [5] is used for alignment; on the other hand, the text representation of entities can be converted into semantic vectors, and the semantic model can be constructed for disambiguation fusion using support vector machine, deep learning and other methods with the associated entity set as the context [6].

2.3 Knowledge Snapshot

The teaching process from simple to deep in colleges and universities enables students of the same major who are enrolled in different years to have highly similar courses in the same grade, and students of the same major to have similar knowledge needs in the same period. User behavior has a guiding effect on users of the same batch or subsequent batches. Therefore, the author proposes to set the grades of different majors as independent concepts. When the grades of users change, a snapshot is taken of the user entity and its relationship network, and then a new user entity is established using the changed data, and a "snapshot" relationship is established between the new user entity and the snapshot user entity, as shown in Figure 1. This method can solve the problem that the knowledge requirement of zero-data user cannot be deduced.

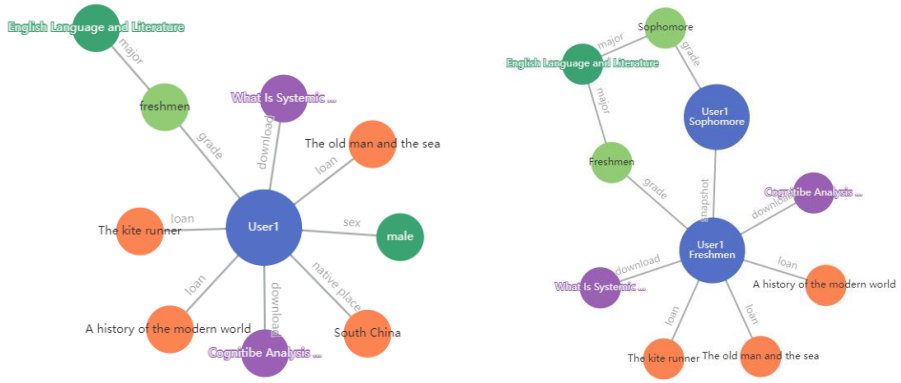

Figure 1 Example of User Upgrade

2.4 Map Optimization

The purpose of graph optimization is to improve the accuracy and efficiency of knowledge derivation by optimizing the entity and relationship network, which includes two aspects. First, it means to eliminate the expired user entity and its relationship, mainly for the users who graduate, quit and retire, and reduce the noise data. The second is to deduce and optimize the relationship between entities, such as deducing the classmate relationship between user entities through the professional grade, deducing the relationship between entities and the author, and finally improving the accuracy of the derivation.

3 Academic resource discovery service based on user graph

Traditional academic resource discovery services aggregate massive academic resource data, and usually provide users with retrieval services in the form of Api interfaces or pages. The process data flow direction is "user → service → user" [7]. Therefore, this paper proposes personalized optimization and reconstruction of process data of discovery service on the basis of traditional academic resource discovery service. According to data input and output, it can be divided into three stages: pre-request expansion, in-process result reconstruction, and post-personalized recommendation. The framework is shown in Figure 2.

3.1 Prior stage

The prior stage is responsible for parsing the user's query input. The author made statistics on the search records of KNOknox in the off-campus access system of digital resources in the library, and found that the average search times of each session were 3.18 times, among which more than 18.5% of the records were fuzzy searches by combining multiple keywords with Spaces, and the users converged the result set by adjusting the query requests several times. In the search service of CNKI, after the first word is entered, CNKI will prompt 1-20 related words or sentences containing the first word, but the content of the prompt does not consider the individual characteristics of the user, and lacks the prompt of keyword combination. To solve the above problems, this paper proposes to transform the query request into the request entity, locate the user and the request entity in the user graph, and use the triangular correlation between the user, the request entity and the academic knowledge entity to realize the personalized recommendation of the query request.

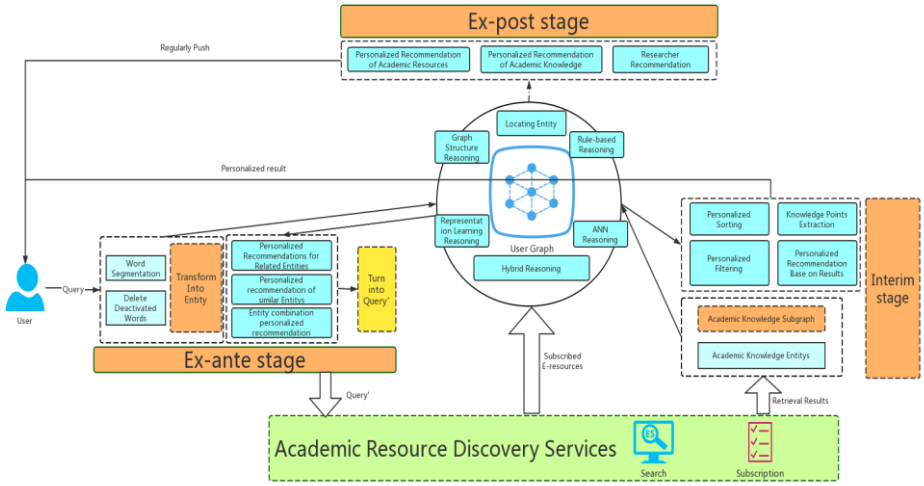

Figure 2 Academic Resource Discovery Service Based on User Graph

3.2 Mid-event stage

The in-process phase refers to the process in which the discovery service returns a collection of academic resource results in response to a query request. In the traditional academic resource discovery service, users still need to converge the result set through secondary retrieval, condition screening and sorting, and the larger the set size, the higher the user behavior cost. Secondly, from the session of digital resource access, it is found that some academic resource catalogue information downloaded by users does not contain search keywords, that is, users obtain resources through data association such as quotation, author and subject term. In the intermediate stage, the result set returned by the discovery service can be converted into academic knowledge and added to the user graph, the correlation between academic resources and users can be used to converge the result set, and related resources can be recommended based on the path relationship between academic resource entities, so as to reduce the cost of user behavior while taking into account the precision and diversity needs of users.

3.3 Post stage

In the post stage, the instant retrieval service scenario is extended to the full-time knowledge service scenario to improve the degree of meeting the user's knowledge needs. At present, the post service of the mainstream discovery system is subscription mode, that is, users manually set the source, keywords, disciplines, authors and other conditions, and push the new academic resources when they meet the requirements. Subscription service requires users to have a certain understanding of themselves and the subject field, and is not suitable for zero-basis users. Limited subscription conditions that need to be set manually are difficult to accurately meet users' dynamic knowledge needs. In the post-event stage, we first subscribe to all academic resources, convert academic resources into entities and add them to the user graph. Then, we take users as the starting point to collect highly relevant entity sets as push objects. Each push will dynamically deduce user needs. For example, hot topic words in the field, authoritative experts, school collaborators, etc.

4 Experiment and discussion based on CNKI

In order to verify the feasibility of user knowledge graph in personalized academic resource discovery service, this paper conducts an experiment using the data accessed by teachers and students from the International Business English School of the university where the author works in the 2022-2023 academic year from the digital resources of CNKI. CNKI is the world's largest provider of Chinese academic resources and plays an extremely important role in academic research in Chinese universities. The coding environment of the experiment was Python3.8, and the knowledge graph construction tool was Neo4j. The experimental data includes postgraduate students, undergraduates and teachers, and the total number of users is 2442. The time range of behavioral data is from June 2022 to May 2023, including 52,252 retrieval behaviors and 28,406 literature download behaviors.

4.1 User map construction

Table 2 lists the extraction of user knowledge. In the knowledge fusion stage, rule dictionaries are established based on the school's management files to align user information entities. Secondly, for factual entities such as authors, publishers and sources of academic resources, Baidu Encyclopedia is used as a third-party knowledge base to establish rule dictionaries for fusion. Finally, semantic entities such as subject words and keywords are targeted. Subject classification of literature and entity co-occurrence rate were used as the basis of fusion. In the phase of graph optimization, 1082 user entities without any behavior records were cut out, and 2762 academic resource entities with missing catalog information and incomplete attribute description were eliminated. Finally, the relationship between users was optimized based on the semester of professional grade. The obtained entity statistics are shown in Table 2.

Table 2 Statistical table of entity data

User related entity		Book related entity		Digital resource entity	
Entity type	Quantity	Entity type	Quantity	Entity type	Quantit
User	1360	Books	3086	docu	25644
Major	5	author	4417	author	31661
College	1	publisher	509	Author Affiliation	8074

Identity Type	3	Subject term	4683	keyword	50753
Grade	7	Chinese library classification number	2138	source	436

4.2 Entity correlation inference

In the experiment, Dijkstra's shortest path algorithm was adopted as the correlation derivation algorithm, which was proposed by Dutch computer scientist Dijkstra in 1959. The algorithm starts from the starting point and uses the strategy of greedy algorithm to find the shortest path to the target node. In order to take into account the influence of the number of paths on the correlation, and take the number of shortest paths as the calculation basis, the algorithm is shown in formula 1.

$$R(x,y) = \frac{n}{D(x,y)} \quad (1)$$

Where x and y are the entities that need to calculate the correlation, are the shortest paths between x and y, and n is the number of shortest paths.

4.3 Personalized query request recommendation experiment

In the experiment, the user query request is intercepted as the original input, and the recommendation algorithm for personalized query request is shown in Table 3. Firstly, the results show that 27.9% of recommendation sets contain complete search requests. Secondly, the experiment compares the query recommendation results returned by the personalized recommendation set and the Knox. Taking a user's query request "network language" as an example, the first three words "network use" are intercepted as input. Table 3 lists the query recommendation returned by Knox and the user knowledge graph respectively. However, semantically, the recommendation result set returned by the user knowledge graph has a higher probability of being close to the user's knowledge requirements. The experiment analyzed the data with large semantic difference between the recommendation result and the complete query request, and found that the main reason for the error was that the user knowledge graph did not contain relevant academic knowledge entities and the shortest path was long. Therefore, in addition to strengthening the construction of academic knowledge, the minimum threshold could also be set to limit the recommendation.

Table 3 Tips for search of Knownet and user graph

Tips returned by CNKI	Tips returned by the user graph
Wang luo yong hu	
Wang luo yong yu	
Wang luo yong hu xin xi	
Wang luo yong hu shu	**Wang luo yong yu**
Wang luo yong hu xing wei fen xi	Wang luo yong yu fan yi
Wang luo yong hu xu qiu	Wang luo yong yu han yi
Wang luo yong hu ti yan	Wang luo yong yu te zheng
Wang luo yong hu xing qu	
Wang luo yong hu guan li	
Wang luo yong hu ying xiang li	

4.4 Search result set personalized reconstruction experiment

The experiment of personalized reconstruction of search result set adopts simulation method and comparison method, that is, the original search result set is obtained by imitating the search behavior of users in the Knownet, and the search result set is personalized reconstructed by using the user knowledge graph. Finally, the number of steps for users to download the target academic resources in the original result set and the personalized result set is compared. According to the experimental results, the user knowledge graph reconstructed more than 80% of the search result sets, and in 62% of the reconstructed result sets, the number of steps for the user to download the target literature was reduced. Taking a user's query request "network terms" as an example, the user downloaded 29 literatures and clicked 41 times on the network. After personalized reconstruction of the initial search result set, it only takes 36 clicks to download the same document, minus 29 clicks of the "download" button, saving about 40% of the behavior. Table 4 reveals the comparison of the number of steps taken by some users to download literature in the original search result set and the personalized result set. Through the analysis of the sorting path, it is found that the historical behavior of users and the sharing of academic knowledge are still the main basis for the correlation between computing resources and users, followed by the behavior of similar users. Through the analysis of the data of reconstruction failure or poor effect, the main focus is on the case of large professional span or high degree of novelty. The experimental results show that the user knowledge graph can reduce the cost of user behavior by reconstructing the original result set and help to get the target resource faster.

Table 4 Simulated retrieval download

Search Term	Number of Articles Downloaded	CNKI Simulation Steps	User Graph Steps
Ying shi fan yi	1	3	1
Dian shi zhi bo	4	6	5
Yu liao ku zai ti	12	16	14
Kua wen hua shang wu gou tong	3	9	4
Cao zong lun	5	8	5

4.5 Personalized academic resource referral experiment

The personalized academic resource recommendation experiment firstly grabs the recent literature published by the mainstream journals of the subject from the knowledge network and adds it to the user knowledge graph, then takes the user as the starting point to obtain the academic resource set with high relevance for recommendation. Take the user who researches "network language" as an example, the user downloaded 45 articles from January to March 2023, involving 38 journals, and the related journals published about 1200 articles in April 2023. The 1200 articles were added to the user knowledge graph, and 75 academic articles were finally recommended to the user. It involves translation, brand, network language and other fields. Judging from the overall experimental results, both old and new literatures are included in the recommendation results set, and the correlation of old literatures is greater than that of new literatures, so it is necessary to consider the publication time and other factors for weight allocation. Secondly, from the perspective of shortest path statistics, The recommendation basis can be divided into "User →

academic resources → academic knowledge → academic resources", "User → academic resources → User → academic resources", "User → professional grade → User → academic resources", in which the user's personal behavior is still the main basis for recommendation. For users whose historical behavior spans multiple disciplines, the results of recommendation are scattered and the effect is not good. Finally, there is still the problem that new academic resources become information islands after joining the atlas, and the correlation cannot be calculated.

4.6 Discussion and development

From the experimental process and results, user knowledge graph can effectively organize user data and optimize the process data of academic resource discovery service, but the experimental results still reflect some problems to be improved:

(1) Academic resource information island problem. When academic resources are added to the graph as independent knowledge entities, it will be impossible to establish relationships with other entities, especially for interdisciplinary frontier knowledge, which will reduce the recommendation effect at each stage. In addition to increasing the volume of academic knowledge, it is still necessary to study the relationship building methods at the semantic level.

(2) Research on correlation derivation methods among different feature entities. The feature dimensions of users and entities related to academic resources are quite different, and whether the correlation derivation between different types of entities can be applied with differentiated methods to reduce errors and improve results needs to be further studied.

(3) Weight calculation of user behavior relationship. Not all user behaviors have continuity, that is, users may study the same topic in different periods or unrelated topics in the same period, and it is necessary to study whether it is possible to dynamically adjust the weight of behavioral relationships through predictive models and preconditions in order to accurately analyze user needs.

5 Conclusion

In the digital age, the organization of user data is one of the important foundational tasks for university libraries to implement intelligent and personalized services. Knowledge graphs have inherent advantages in information organization and relationship discovery, making them highly compatible with the structured and semi-structured user and academic resource data in libraries. This article focuses on university library users and proposes a process for constructing a user graph in university libraries. It explores the integration of user graphs into different stages of academic resource discovery services and demonstrates the effectiveness of user graphs through experiments. The aim is to provide insights for the research and application of knowledge graphs in organizing and utilizing user data in university libraries.

Acknowledgments

This article thanks Youth Program of Philosophy and Social Sciences of Guangdong Province, (Grant No.GD20YTS01; No.GD21YYJ01;); Youth Program of Guangdong Basic and Applied Basic Research Fund Committee-Regional Joint Fund, (Grant No.2022A1515110873);

Characteristic Innovation Project of Guangdong Provincial Department of Education (Social Sciences), (Grant No. 2022WTSCX014)

References

[1] Li Huifang, Meng Xiangbao. Review and Analysis of Research and Practice on Library Resource Discovery Systems at Home and Abroad in the Past Decade[J]. Library and Information Service, 2020, (6): 120-129.

[2] Zhang Yushu. Research on Search Engine Technology Based on Knowledge Graph[J]. Information Technology and Informatization, 2020, (9): 29-31.

[3] Zhang QH, Sun ZQ, Hu W, et al. Multi-view knowledge graph embedding for entity alignment. Proceedings of the 28th International Joint Conference on Artificial Intelligence(IJCAI). Macao, 2019: 5429.

[4] Xie RB, Liu Z, Jia J, et al. (2016) Representation Learning of Knowledge Graphs with Entity Descriptions. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence, AAAI, Menlo Park, 2659-2665.

[5] Alokaili A, Menai M. SVM ensembles for named entity disambiguation[J]. Computing: Archives for informatics and numerical computation, 2020, 102(4): 1051-1076.

[6] Zhu Ganggao, Iglesias, et al. Exploiting semantic similarity for named entity disambiguation in knowledge graphs[J]. Expert Systems with Applications, 2018, 101(Jul.): 8-24.

[7] Li Huifang, Meng Xiangbao. Review and Analysis of Research and Practice on Library Resource Discovery Systems at Home and Abroad in the Past Decade[J]. Library and Information Service, 2020, 64(06): 120-129.