Integration of Fragmented Knowledge Based on Knowledge Graph

Xiaojun Chen[a*], Liu Yuan[b]

College of Computer Science, Shaanxi Normal University, China

[a*]Xiaojunchen_xj@163.com, [b]yuanliu@snnu.edu.cn

Abstract. With the widespread adoption of user-generated content (UGC) platforms, the influx of fragmented knowledge makes it increasingly complex to integrate and utilize this knowledge across platforms. This study proposes an innovative approach based on knowledge graph and PageRank algorithm to effectively integrate fragmented knowledge in different UGC platforms. In this paper, fragmented knowledge from different platforms is collected. After data preprocessing, PageRank innovation algorithm is used to calculate the importance score of each knowledge node, and the relevance and importance of nodes are taken as the basis of integration, and it is organically organized into a unified knowledge graph. Nodes in the graph represent knowledge elements on different UGC platforms, while edges represent the relationships between them, forming a hierarchical integrated knowledge graph. Experiments show that this method can not only significantly improve the quality of integrated knowledge, but also effectively solve the problem of information fragmentation among different UGC platforms. This research provides an innovative solution for cross-platform fragmented knowledge integration, so as to help learners make better use of fragmented knowledge to improve learning effect, provide better learning resources and guidance for educators and learners. It is expected to be widely used in the field of knowledge management and integration.

Keywords: Knowledge graph; PageRank innovation algorithm; Fragmented knowledge; Knowledge integration

1 Introduction

In today's information age, the cyberspace is full of huge and fragmented knowledge[1]. Fragmented knowledge is usually scattered and fragmented knowledge obtained from different sources and channels, which is distributed in different network platforms, subject areas and contexts, forming a complex and intricate knowledge network. Users are often faced with the challenge that fragmented knowledge is difficult to integrate when they pursue comprehensive understanding and access to information. Although the current search engine has made remarkable progress in providing information retrieval[2], it mainly relies on keyword matching, and users are often faced with the situation of

Y. Kuang et al. (eds.), *Proceedings of the 2024 5th International Conference on Education, Knowledge and Information Management (ICEKIM 2024)*, Atlantis Highlights in Computer Sciences 22,
https://doi.org/10.2991/978-94-6463-502-7_57

scattered and fragmented information, and are often unable to effectively understand and integrate cross-domain and cross-platform knowledge fragments.

In this context, the emergence of knowledge graph provides a new way to integrate fragmented knowledge. With its structured representation of entities and relationships, knowledge graph can better express the structured correlation between knowledge and connect knowledge nodes in different domains and platforms. At the same time, PageRank algorithm, as a classic network analysis tool, is widely used in web page ranking. However, in the knowledge graph, the traditional PageRank algorithm may not fully consider the structural association between nodes, thus affecting the overall knowledge integration effect.

There are still some limitations in the intelligence and accuracy of integrating fragmented knowledge. Therefore, this study aims to deeply explore the combination of knowledge graph and PageRank algorithm, solve the problem of fragmented knowledge integration through innovative methods, and provide users with more comprehensive and in-depth knowledge acquisition experience. Through this research, we expect to make a positive contribution to solving the problem of information fragmentation, promoting the progress in the field of knowledge integration and better meeting the knowledge needs of users in the information age.

2 Related work

2.1 Fragmented Knowledge Integration Methods

Fragmented knowledge comes from a wide range of sources and the content quality is not the same. How to collect, integrate and use fragmented knowledge to realize the ordering and aggregation of knowledge is an urgent problem to be solved at present. Sharma et al. realized efficient knowledge organization on the semantic Web by changing the ontology, and improved the results of Web mining by using machine understanding to represent the structure of web documents[3]. Zheng et al. built a faceted hierarchy based on the correlation between facets through the knowledge forest, and organized the facets into a tree-like structure according to the correlation strength to integrate fragmented knowledge[4]. Liang et al. analyzed the characteristics of fragmented learning behavior, reorganized knowledge in online education according to learners' individual learning needs, and helped and guided learners to make full use of fragmented time to obtain accurate and meaningful knowledge content[5]. In order to better understand the methods used for fragmented knowledge integration, we made a feature comparison between traditional knowledge integration methods and knowledge graph-based fragmented knowledge integration methods, as shown in Table 1.

Table 1. Comparison between traditional knowledge integration and fragmented knowledge integration based on knowledge graph

Feature	Traditional integration	Knowledge graph integration
Data sources	From professional databases	From multiple UGC platforms
Data treating	Manual induction	Automatic scraping

Data structure	Adopt planar structure	Graphical structure
Relevance	Human judgment and rules	Automatic association rule
Visualization	Limited visualization	Rich graphical display
Application area	Specific field or industry	Cross-domain integration

2.2 Application of Knowledge Graph

Knowledge graph is applied to various scenarios. Chandak et al. used knowledge graph to realize precision medicine and connect fragmented biomedical knowledge with patient-level health information[6]. For a given disease, information from these organizational scales is dispersed across different publications and data repositories, and grid-based relationships are developed among these sources for precision medicine research. Due to differences in domain and understanding, Chinese grottoes with semantic 3D modeling pose great challenges to individuals lacking professional knowledge of cultural heritage. In order to overcome these obstacles, Yang et al. proposed a knowledge graph representation to provide explicit knowledge for participants in different stages of semantic 3D modeling of Chinese grottoes[7]. Railway loop optimization is a complex process that extensively uses human knowledge and experience, which is difficult to be recognized by computer. In order to solve the above problems, Pu et al. proposed the earliest known knowledge graph modeling method for railway alignment optimization, designed a hierarchical classification semantic network modeling method for railway alignment design knowledge, and built the knowledge graph for railway alignment design on this basis[8].

2.3 PageRank Algorithm and Its Application

In recent years, Was et al. have proposed more than a hundred centrality measures, each evaluating the location of nodes in a network from a different perspective, investigating the fundamental problem of identifying the most important nodes in a network, and providing the first axiomatic characterization of the general form of PageRank[9]. How to identify important nodes in multi-layer networks is still an unsolved problem in network science. In the past few decades, Lv et al. have defined various centrality methods from different perspectives to find influential nodes in multi-layer networks, and weighted the shortest path between any two nodes in all network layers[10].

3 Fragmented Knowledge Structure

The integration of fragmented knowledge is one of the important challenges in the field of knowledge management. In this section, we will analyze the structure of the knowledge graph, which is the core of integrating fragmented knowledge. In the knowledge structure diagram in Fig. 1, we can see that the knowledge unit structure is described as G = (C, S, N, D, W), where G represents the knowledge graph, C represents the domain scope of the graph, S represents the knowledge source, N represents the

knowledge node, W represents the edge weight, and D represents the degree of the knowledge node.

C = {C_1, C_2, ..., C_n} is a knowledge domain, representing a collection of source domains of fragmented knowledge, including fragments of knowledge in n different domains. S = {S_1, S_2, ..., S_z} is a collection of sources. Knowledge points in the same field may come from different UGC user generation platforms, and there are z sources of knowledge points. We need to collect fragmented knowledge in the same field from different sources. N = {N_1, N_2, ..., N_m} represents the knowledge point set, with a total of m knowledge nodes, and the more knowledge nodes, the more fragmented knowledge. W = {W_1, W_2, ..., W_k} represents the set of edge weights of knowledge nodes, the larger the edge weights, the more important the relationship between two knowledge nodes is in the whole knowledge structure system, W_k represents the weight of the KTH knowledge node. D = {D_1, D_2, ..., D_t} represents the degree set of knowledge nodes, the greater the degree of a knowledge node, the closer the association with other nodes, dt represents the degree of the t node.

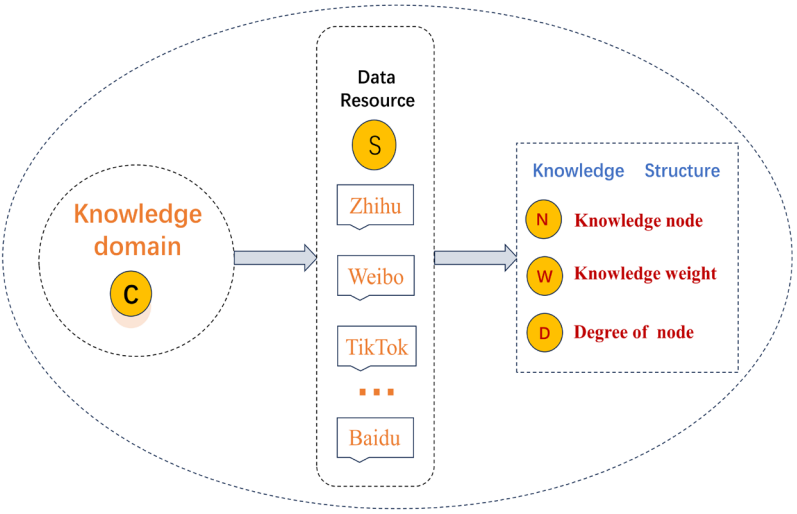

Fig. 1. Knowledge structure diagram

4 Knowledge Graph Integrates the Improved PageRank Algorithm of Fragmented Knowledge

The ordinary PageRank algorithm has certain limitations, which cannot clearly show the importance of different knowledge nodes, thus affecting the connection effect of different nodes in the knowledge graph. It needs to be improved. The specific formula after improvement is shown in (1).

$$\mathrm{PR}(V_i) = \frac{1-q}{N} + q\sum_{i\in N(j)} \frac{\mathrm{TF}(i)\cdot \mathrm{PR}(V_j)\cdot \mathrm{Weight}(v_j)}{\mathrm{Degree}(v_j)\cdot \sum_k TF(k)} \tag{1}$$

In formula (1), Where q is the damping factor, typically set to 0.85, and N represents the total number of nodes. Degree (V_j) is the total Degree of V_j and Weight (V_j) represents the weight of the edge (V_i, V_j). We replace the output of the calculated vertices in the original algorithm with the sum of degrees of the vertices Degree(V_j), which is the sum of the degrees of the vertices calculated. In this way, the directionality is ignored and the algorithm is applied to the undirected graph. Then multiply the Weight of the edges (V_i, V_j) by Weight (V_j), the greater the weight between the edges, the more important the relationship between the entities. TF(i) represents the word frequency of knowledge node i, $\sum_k TF(k)$ represents the sum of the word frequency of all knowledge nodes. On the basis of the original algorithm, TF can better reflect the importance of keywords in the text, so as to be more accurate in the processing of text data and further deepen the correlation degree. We derive and expand formula (1) and combine TF(i), Weight(V_j), Degree(V_j), $\sum_k TF(k)$ from the algorithm to obtain θ_{ij}, presenting it as a term in the derived formula (2).

$$\mathrm{PR}(V_i) = \begin{bmatrix} \mathrm{PR}(V_1) \\ \mathrm{PR}(V_2) \\ \vdots \\ \mathrm{PR}(V_n) \end{bmatrix} = \frac{1-q}{N} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + q \begin{bmatrix} \theta_{11} & \theta_{12} & \cdots & \theta_{1n} \\ \theta_{21} & \theta_{22} & \cdots & \theta_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ \theta_{n1} & \theta_{n2} & \cdots & \theta_{nn} \end{bmatrix} \begin{bmatrix} \mathrm{PR}(V_1) \\ \mathrm{PR}(V_2) \\ \vdots \\ \mathrm{PR}(V_n) \end{bmatrix} \tag{2}$$

In formula (2), Where PR(V_i) is the new PageRank value of node V_i. The specific expression of θ_{ij} is given by formula(3). When V_i and V_j are not connected, $\theta_{ij} = 0$. The larger the word frequency of node Vi, the larger θ_{ij} and the larger the weighted degree of node V_j, the larger θ_{ij}.

$$\theta_{ij} = \frac{\mathrm{TF}(i) \cdot \mathrm{Weight}(v_j)}{\mathrm{Degree}(v_j) \cdot \sum_k TF(k)} \tag{3}$$

First we initialize each node with a PageRank score of 1.0, which means that each node has the same initial importance. Then the above calculation steps are repeated to iteratively calculate the PageRank score of each node. In the calculation process, the PageRank score of the target node is updated by considering the PageRank score of the neighbor nodes and the number of their neighbor nodes. After each iteration, check whether the PageRank value converges. A convergence threshold is set, and the algorithm stops iterating when the change in PageRank value is less than the threshold. Once the iteration is complete, the PageRank value of each node converges to the final result, and the PageRank value of the node is updated through multiple iterations, Finally, the final sorting result is generated according to the importance of nodes.

The main steps of knowledge graph construction are as follows. To begin with, we preprocess the collected fragmented knowledge, add the source node, relation node and target node of each knowledge fragment to the knowledge graph by using Gephi software. Then, we apply the improved PageRank algorithm to the knowledge graph, The attributes of knowledge nodes in the graph are updated according to the ranking results of PageRank values of different knowledge nodes after iterative calculation, and finally output the knowledge graph integrated by the algorithm. The graph contains all the nodes and the relationships between them, which can help us better understand and use the fragmented knowledge.

Our pseudocode for integrating fragmented knowledge using knowledge graph with improved PageRank algorithm is shown in Table 2.

Table 2. Pseudocode for integrating fragmented knowledge with PageRank algorithm

Algorithm: Integration of Fragmented Knowledge using PageRank
Input: Crawled knowledge;
Output: Integrated knowledge graph;
Step 1: Construct knowledge graph;
Step 2: Initialize PageRank scores;
Step 3: Iteratively calculate PageRank scores until convergence;
Step 4: Update PageRank scores;
Step 5: Get nodes sorted by PageRank scores;
Return Integrated knowledge graph;

5 Experiment and Results

5.1 Experimental Data Source and Knowledge Graph Integration Process

In this experiment, we used the Octopus collector to collect knowledge points related to data structure from Tik Tok, Weibo, Zhi hu, Post Bar and classroom PPT as research cases. After data collection, word segmentation and other data preprocessing operations, 161 knowledge nodes and 154 knowledge data of different node relationships were obtained. These data are saved in .CSV format, and then the improved PageRank algorithm is used to mine the entity relationship. Finally, the relationship between knowledge elements is displayed by knowledge graph. The algorithm first constructs the knowledge graph and integrates the fragmented knowledge captured from different UGC platforms into a hierarchical graph, where nodes represent knowledge points and edges represent the correlation between knowledge points. Then, by using edge weights and vertex degrees, the improved PageRank algorithm is used to calculate the scores of nodes in the graph, consider the relationship between nodes and their neighbors, and determine the importance of nodes. In the iterative process, the PageRank score is calculated by updating and converging, and the relative importance of each node is finally obtained. By sorting nodes according to PageRank score, the effect of the original knowledge graph is improved.

5.2 Integration Results of Knowledge Node PageRank Value Distribution Table and Knowledge Graph

We show the knowledge nodes of the Top6 PageRank values in the experiment as well as the degrees and weights of the nodes in Table 3. We present the integrated knowledge graph in Fig. 2, In the graph, we can see that knowledge points from different sources are integrated, and knowledge points from different sources are displayed in different colors in the graph. In the graph, we can see that different knowledge nodes have different sizes, and the importance of different knowledge is demonstrated by the size of knowledge nodes. The larger the node is, the more important the knowledge point is. When learning and applying related knowledge, we can learn different knowledge

points in turn according to their importance, so as to improve the learning efficiency and make fragmented knowledge easier to understand and apply. We observe that the integrated knowledge graph is significantly expanded in scale, which verifies the stability and effectiveness of the algorithm.

Table 3. PageRank value, degree and weighting degree table of knowledge node

Node	PageRank Value	Degree	Weighting degree
DS	0.168895	5	20.0
Tik Tok	0.157894	4	18.0
Zhi hu	0.128361	4	13.0
PPT	0.148981	5	17.0
Post Bar	0.142281	5	16.0
Weibo	0.126443	3	10.0

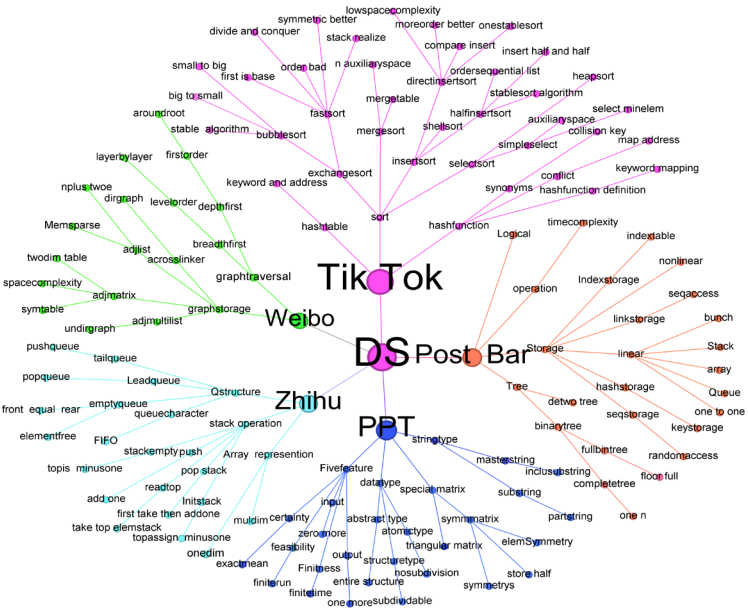

Fig. 2. Knowledge map

In our study, we collected fragmented knowledge from different UGC platforms. Although our data came from multiple platforms, there may still be insufficient data in some fields or topics, which may lead to the biased integration results in some aspects. Future studies may consider choosing more extensive data sources to improve the comprehensiveness of knowledge integration.

6 Conclusions

In this study, we explore the method of integrating fragmented knowledge by combining knowledge graph and PageRank innovation algorithm, deepen and fill the knowledge integration domain method, and provide an effective way to organize and utilize multiple fragmented knowledge. In the future, we will continue to work on improving integration methods to further improve the accuracy and efficiency of integration. At the same time, we will explore more integration methods and apply integration methods to reasoning, prediction and recommendation, so as to apply the integrated knowledge graph to a wider range of application scenarios.

References

1. Liang, K., Wang, C., Zhang, Y., Zou, W. (2018) Knowledge Aggregation and Intelligent Guidance for Fragmented Learning Knowledge Aggregation and Intellige. Procedia Computer Science, 131: 656–664. doi:10.1016/j.procs.2018.04.309.
2. Shakeri, M., Sadeghi-Niaraki, A., Choi, S.-M., AbuHmed, T. (2022) AR Search Engine: Semantic Information Retrieval for Augmented Reality Domain. Sustainability, 14: 15681-15697. doi:10.3390/su142315681.
3. Sharma, A., and Jain, S. (2022) Altering OWL Ontologies for Efficient Knowledge Organization on the Semantic Web. International Journal of Information System Modeling and Design, 13: 1-16. doi:10.4018/ijismd.313431.
4. Zheng, Q., Liu, J., Zeng, H., Guo, Z., Wu, B., Wei, B. (2021) Knowledge forest: a novel model to organize knowledge fragments. Science China Information Sciences, 64: 179101–179103. doi:10.1007/s11432-018-9940-0.
5. Liang, K., Zhai, J., Ren, Y., Zhang, Y., Wang, Y. (2022) Fragmented Knowledge Clustering method based on SOM. International Journal of System Assurance Engineering and Management, 14: 188-195. doi:10.1007/s13198-021-01504-1.
6. Chandak, P., Huang, K., Zitnik, M. (2023) Building a knowledge graph to enable precision medicine. Scientific Data, 10: 67-83. doi:10.1038/s41597-023-01960-3.
7. Yang, S., and Hou, M. (2023) Knowledge graph representation method for semantic 3D modeling of Chinese grottoes. Heritage Science, 11: 266-292. doi:10.1186/s40494-023-01084-2.
8. Pu, H., Hu, T., Song, T., Schonfeld, P., Wan, X., Li, W., Peng, L. (2024) Modeling and application of a customized knowledge graph for railway alignment optimization. Expert Systems with Applications, 244: 122999-123013. doi:10.1016/j.eswa.2023.122999.
9. Wąs, T., and Skibski, O. (2023) Axiomatic characterization of PageRank. Artificial Intelligence, 318: 103900-103938. doi:10.1016/j.artint.2023.103900.
10. Lv, L., Zhang, T., Hu, P., Bardou, D., Niu, S., Zheng, Z., Yu, G., Wu, H. (2024) An improved gravity centrality for finding important nodes in multi-layer networks based on multi-PageRank. Expert Systems with Applications, 238: 122171-122185. doi:10.1016/j.eswa.2023.122171.