Academic New Media Service Method Based on Knowledge Graph

Chuzhen Zhou*

*zhou_chuzhen@163.com

Liaoning Technical University, Longwan South Avenue, Xingcheng City, China

Abstract: [Purpose/Significance] The academic information service is an important basis for the development of current academic research. Traditional academic information services focus on the discovery and mining of information in the process of service provision, while ignoring that their core focus should be on user needs. New media services are communication media and platforms that have been widely used in recent years. They are characterized by ubiquity, customization, and user participation. [Method/Process] In order to further improve the quality of academic information services with new media services, we propose to use interknowledge relationships to map intellectual entities, and then achieve the purpose of customized recommendations based on user preferences and needs. [Results/Conclusions] The experimental results on multiple data sets show that the proposed method is able to more accurately and quickly complete customized recommendations for the academic services provided to the users.

Keywords: knowledge graph, academic new media, information service, customized recommendation

1 INTRODUCTION

With the rise of new media services, the dissemination and customization of information have been greatly improved. Therefore, in the fields of library intelligence, government affairs, rural agriculture, national meteorology, etc., there has been a gradual shift to the use of new media to carry out information services. For the academic field that requires high standards in terms of user level, information quality, and communication efficiency, the application of new media platforms to achieve academic information services is also a general trend. Both new media services and academic information services are "people-oriented" services with the goal of meeting user needs. This article divides user needs from the perspective of online marketing and explains their evolutionary rules, academic information service dynamics, information flow mechanisms, and new media service modes, hoping to provide new research ideas for academic new media information services.

Academic new media is an emerging term. It can be seen as an interactive and integrated media form and platform created in the process of the dissemination and interaction of academic information, and is based on digital technology, network technology, and other modern information technology or communication technology [6]. As a medium and a platform, it is responsible for the communication between users and the system, emphasizing the efficiency and quality of information interaction with a view to achieving user satisfaction. Sun et al [10]

EMIS 2023, February 24-26, Chengdu, People's Republic of China

DOI 10.4108/eai.24-2-2023.2330612

believe that information service is a user service and user information activity based on the relationship between information and users, and it is a service activity that promotes and guarantees users' perception, absorption and utilization of information. The needs of users in media services are mainly embodied in precise targeting, personalization, and emphasis on experience and service. As a development orientation for information services, academic information services primarily need to meet the needs of users, and then focus on studying the collection, organization, processing, and analysis of information resources in information systems.

Academic new media information services can be seen as a combination of new academic media and academic information services, covering the entire process of academic services, including research on user information behavior, academic information dissemination, and academic information work within information systems. Therefore, this paper defines academic new media information services as the information services conducted on the platform of academic new media through the search, sorting out, processing and analysis of the academic information, which are provided to academic users such as students, teachers, researchers or R&D personnel to meet their needs by the academic institutions, academic journals, digital media companies and other service providers.

With the development of the technologies such as big data, the Internet, and mobile communications, scholars have conducted in-depth research on academic new media information services from different perspectives. From the perspective of academic new media user behavior, Zhang et al. [13] proposed that the exploration of mobile terminal users' information behavior and its influencing factors should be strengthened through the analysis and utilization of massive data generated by user behaviors. From the perspective of academic new media communication methods, Ren [9] believes that academic journals have flaws in editing and publishing, mindset, content dissemination methods, and cross-media management, and proposed improvement measures for in-depth and comprehensive media integration. From the perspective of the academic information service model, Qu et al. [8] discovered that the existing digital libraries have barriers to information exchange by investigating the current state of academic information services in the college digital library community, and proposed a new model for the college digital library community based on information exchange needs. From the perspective of developing and designing academic information service platforms, Fu et al. [1] designed scientific research information service platforms for different user levels based on the needs of university research information users.

The above studies cover the entire process from user demand through user information services to demand fulfillment. It also explains information services and academic new media from the application of specific academic new media information services to widely existing user behavior and user needs. As can be seen from the concept of academic new media information services, academic new media information services include user-centered information resource integration functions emphasized by information services, as well as user-oriented information dissemination and platform development and design processes focusing on new media. Therefore, the research on academic new media information services should focus on a systematic study of the entire process of academic services, not just a certain process, field, or platform. Based on the above research results, this paper analyzes the academic needs from the perspective of a knowledge graph, analyzes information flow rules guided by the evolution of demand, proposes service models that meet the needs of different academic users, and realizes information dissemination and interaction within the service model..

2 ACADEMIC NEW MEDIA SERVICE METHOD BASED ON KNOWLEDGE GRAPH

In academic new media information services, many scholars have analyzed user needs from many different angles. For example, based on how the needs of university library users are realized, users are divided into visiting users and non-visiting users (Fu 2011); according to library user satisfaction, they are divided into basic needs, expected needs, exciting needs, and undifferentiated needs [11]; according to users' cognitive goals, their needs are divided into understanding needs and creative needs [5]. In this paper, the learning style of users is used as a breakthrough point to analyze customer demand orientation to provide targeted services.

Learning style is a relatively stable state in the learning process, which is a collection of dimensions of the learners' perception, acquisition, processing and understanding of knowledge. Through analysis of image models and applications, we believe that using the Felder-Silverman Learning Style Model (FSLSM) model to construct images [4] is beneficial to improving the user experience of customized resource recommendations.

2.1 Interest keyword entity extraction algorithm

After obtaining the customer's learning style, extracting interest keyword entities as the basis for service recommendations can improve the user's experience and the accuracy of interest recommendations and service recommendations. The first step in extracting keywords of interest is to identify named entities and extract entities from structured or unstructured data. As important carriers of information, entities can be various elements, names of people, place names, or even a concept. The extraction methods range from early extractions through manual operation, string matching, etc., to later extractions through automated methods such as natural language processing. Currently, deep learning is also used in knowledge graphs for named entity recognition.

As a commonly used keyword extraction technique, the TF-IDF (Term Frequency-Inverse Document Frequency) [2] algorithm evaluates the importance of keywords in the text by calculating term frequency (TF) and inverse document frequency (IDF). The frequency of terms corresponds to the frequency with which the term appears in the text. To a certain extent, it can be understood as the importance of a term in the text which is directly related to the number of its occurrences, that is, the higher the term frequency, the more important the term is. The inverse document frequency indicates the frequency at which the term appears in the corpus. Contrary to term frequency, the higher the inverse document frequency, the higher its popularity, and the lower its probability of becoming a keyword. The calculation formula consists of two parts.

$$\mathrm{TF} = \frac{f_{i,j}}{\sum_k f_{k,j}} \tag{1}$$

where $f_{i,j}$ indicates the number of the word w_i appearing in the document d_j, and TF is the weight of w_i in the document d_j.

$$\mathrm{IDF} = \lg \frac{|\mathrm{D}|}{|1 + \{\mathrm{j}: \mathrm{w_i} \in \mathrm{d_j}\}|} \tag{2}$$

where D is the total amount of documents in the corpus,$\{\mathrm{j}: \mathrm{w_i} \in \mathrm{d_j}\}$ indicates the number of documents in the corpus where $\mathrm{w_i}$ appears, TF represents the term frequency, corresponding to the frequency of occurrence of a term in the document, which is introduced in Formula (1) above. The domain of k can be found in the sigma notation.

Combining Formula (1) and Formula (2), the final TF-IDF is calculated as follows:

$$\mathrm{TF} - \mathrm{IDF}(\mathrm{w_i}) = \mathrm{TF}(\mathrm{w_i}) * \mathrm{IDF}(\mathrm{w_i}) = \frac{\mathrm{f_{i,j}}}{\sum_k \mathrm{f_{k,j}}} * \lg \frac{|\mathrm{D}|}{|1 + \{\mathrm{j}: \mathrm{w_i} \in \mathrm{d_j}\}|} \tag{3}$$

Through the processing and analysis of each word in the identification subject, the TF-IDF value of each word is obtained, which is used as the basis for keyword selection in turn (TF-IDF is often used to mine the keywords in articles, and the mean value of TF-IDF represents the appearance of all keywords). The basis for the successful execution of the TF-IDF algorithm is the accurate recognition of each entity word. Recurrent neural networks (RNN) have achieved good results in scenarios such as classification and entity labeling. The long-term short-term memory network model (LSTM), as a variant of RNN, improves the ability to obtain document information by increasing time series relationships. Using the gated concept, the three parts of the forgotten gate, input gate, output gate, and memory cells are defined. The process of memory storage and transfer was simulated to achieve selective long-term memory, effectively mitigating the problems related to gradient disappearance and gradient explosion. The hidden layer portion is also known as the LSTM unit, as shown in Figure 1.

The forgotten gate determines whether the information in the memory cell needs to be forgotten. The input at the current moment, the state of the hidden layer at the previous moment, and the state of the memory cell at the previous moment are input as data, and output after activating the Sigmod function with $\mathrm{f_t} = 1$ or 0. The former indicates complete retention, while the latter indicates complete oblivion. Its forward propagation formula is:

$$\mathrm{f_t} = \sigma(\mathrm{W_{xf}x_t} + \mathrm{W_{hf}h_{t-1}} + \mathrm{W_{cf}c_{t-1}} + \mathrm{b_f}) \tag{4}$$

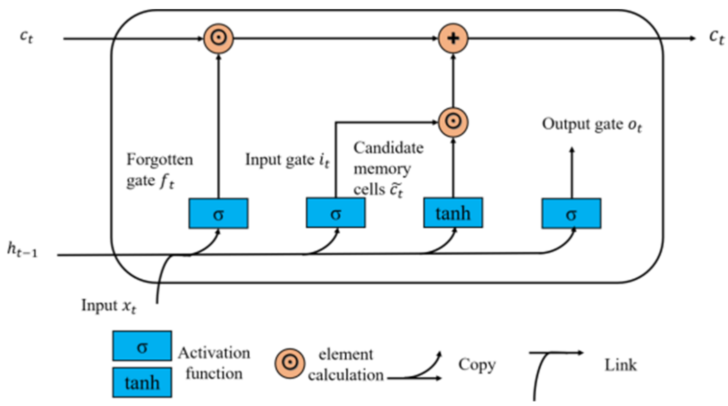

Figure 1: LSTM unit.

The input gate is used to determine whether to retain information. It uses the same activation function, and its propagation formula is as follows:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \tag{5}$$

where i_t= 0 or 1 $i_t = 0$ means discarding the current information, otherwise keep it. The information to be added depends on the input at the current time and the state of the hidden layer at the previous moment. The information to be added is denoted as z_t, which is defined as follows:

$$z_t = \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{6}$$

Candidate memory cells are used to indicate stored information, and at the same time update memory cells according to whether past information and new information added need to be retained, which is expressed as follows:

$$c_t = f_t c_{t-1} + i_t z_t) \tag{7}$$

The input gate determines the output of the current layer information. For time t, if $o_t = 0$, it means that the current layer does not output, if $o_t = 1$, it means output, the formula is as follows:

$$o_t = \tanh(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \tag{8}$$

The content to be output h_t is determined by the information in the memory cell at time t after being normalized by the hyperbolic tangent function and the output gate. The formula is expressed as:

$$h_t = o_t \tanh(c_t) \tag{9}$$

2.2 Interest Keyword Extraction Based on BERT + Bi-LSTM + CRF

In this section, we use a bidirectional long-term short-term memory network (Bi-LSTM) (Zhou2016) and an extraction model combining Conditional Random Fields (CRF) to extract keywords of interest. First, we generate deep bidirectional semantic features based on BERT, then use the characteristic of Bi-LSTM that preserves the back-and-forth relationships of semantics to encode the upper-level output, and finally decode the coding results using CRF, so as to obtain labeling results with higher accuracy.

2.3 Entity representation of interest keywords entities based on knowledge graphs and TransH models

By abstracting a relationship-oriented hyperplane, TransH (Wang 2014) projects the head and tail entities through the projection vector w onto the spatial hyperplane abstracted by that relationship. Separating entities and relationships makes entities have different representations in different relationships, which mitigates the problems of one-to-many and many-to-many projections.

The head and tail entity vector of the TransH model passes w, and the vector is mapped to the plane where the relationship r is located. The goal is to minimize the sum of the loss function and adjust the error between the first entity vector and the tail entity vector on that plane through d after addition. But due to the uncertainties in the random initialization operation, the training time was too long.

In order to shorten the training time, using the characteristics of the relatively simple node relationships and clear semantic properties in the subject knowledge graph, a pre-trained language model is combined with relational vector initialization during the relationship initialization process, and the three-tuple structure is trained on this basis. Entity relationships are converted into low-dimensional vectors based on semantics, so that the model generates an initialized hyperplane space with a semantic relationship after the relationship is initialized to achieve the goal of bringing close the relative position of the initialization plane and the relative position of the final space. Since the vector translation does not deform, the distance that needs to be adjusted during model training is reduced, and efficiency is improved by reducing the number of iterations of sampling training for each three-tuple group.

2.4 Customized recommendation algorithm based on interest keywords entities

A subject knowledge graph is a structured knowledge base composed of knowledge entities of a certain subject and relationships between various knowledge entities. The entity consists of the associated knowledge points or educational resources. Each entity is distinguished by a globally unique identifier, similar to the classification of a web page through the address area of a web page, and the relationship between entities describes the relationship between knowledge entities and intellectual entities, and between entities and educational resources. Furthermore, the construction of a subject knowledge graph is mainly based on requirements with standard references, such as curriculum standards and subject education rules, and fully takes into account the logical relationships between knowledge entities.

In response to the structural relationship features of the knowledge graph, we selected relationship features and tested them using a weighted importance ranking algorithm. There are differences in the importance of different attribute nodes to knowledge entities and in the structure of the graph. Moreover, in most cases, the importance conveyed by the connection relationship of key attributes is significantly higher than that of ordinary links. At the same time, there is also a relationship between some connections and the user's learning style. The PageRank algorithm cannot be used as a basis for adaptively adjusting the importance transmission of intellectual entity resources according to the user's learning style. To this end, we give different weights to different relationships, so that the KR values can be adjusted at a more granular level when transferred, so as to obtain a ranking of KR values that better meet the needs.

By analyzing the role of each relationship in the connection of node relationships and the actual needs of users, the corresponding relationship weight matrix is obtained. By adjusting the weight matrix, the KR value corresponding to each node in the subject graph can also be changed, and finally, the KR value sorting sequence required by the user will be obtained, which lays the foundation for the customized recommendation for the user's needs.

3 EXPERIMENTAL VERIFICATION AND EXPERIMENTAL ANALYSIS

In this article, we use the MAKG data set to build a knowledge graph to verify the effectiveness of our approach. Specifically, MAKG is a very large RDF format academic data set containing over 8 billion triplets. This graph of academic knowledge was compiled and published by Microsoft Corporation. The data set has obtained an open data-sharing signature license. It can be downloaded, and the data set is automatically updated through Microsoft's services every certain period of time. Entities in the knowledge graph are represented by URIs, which can be linked through open cloud data and other data sources. The statistics on the number of various types of entities included in this data are shown in the following table:

Table 1: Statistics of the main information volume of the MAKG dataset

Entity	Quantity
Publications	238,670,900
Authors	151,355,324
Affiliations	25,767
References	1,635,169,990
Citations with citations contexts	234,337.833
Fields of study	740,460
ORCID IDS	34,863
Conference instances	16,142
Journals	48,942
Paper abstracts	139,227,097
Paper tags	677,389,638
Linked datasets and code	75,091

During the training process for 100 rounds of the training set, both algorithms (TranSH and our method) were randomly sampled. The comparison of the running times expressed by the two interest keyword entities is shown in figure 2:

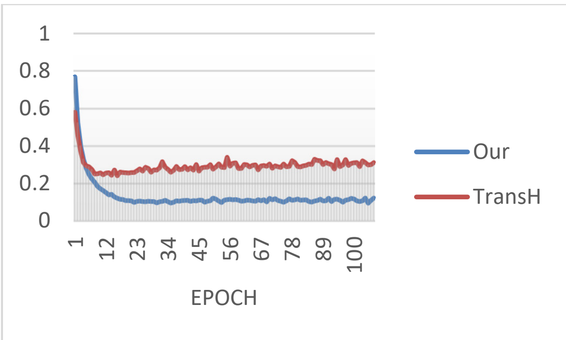

Figure 2: Algorithm time consumption comparison chart.

From the figure, we can see that due to the random initialization of tranSH, the training time for locally sampled data during the training process was extended, and the training took a relatively long time throughout the training process. In the same case, since the algorithm we proposed based on the knowledge graph combined semantic relationships when the relationships were initialized, the results at the time of relationship initialization were closer to reality. It shortened the training time and improved training efficiency.

In addition, we also recorded the loss values of the two methods, as shown in Figure 3. From the chart, we can see that as the number of training rounds increased, the loss values of the two algorithms converged to similar values, thus indicating that the ability of the two to express knowledge is basically the same, thus proving that our method is feasible in representing knowledge.

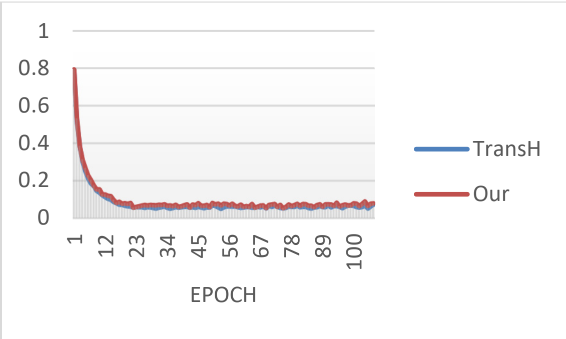

Figure 3: Comparison chart of loss values.

Finally, we compared the performance of the two methods under the three metrics of Recall, Precision, and F1. The experimental results are shown in figure 4. From the figure, we can see that the proposed method is superior to the compared method under the three measurement indicators.

Figure 4: Comparison of Recall, Precision, and F1 values.

4 CONCLUSIONS

Both information services and new media services are the research focuses of academic services today. Although the service focuses are different, they are all developed with user needs as the core. New media services are one of the methods that have been widely used in communication media and platforms in recent years and have received widespread attention from researchers. However, the existing new media service method still has shortcomings in the customized recommendation method. For this reason, this paper proposes a way to map knowledge entities using relationships between knowledge, and then achieve the purpose of customized recommendations based on user preferences and needs. The experimental results on the MAKG data set show that the proposed method can more accurately and quickly complete customized recommendations for academic services for the users..

REFERENCES

[1] Fu K L. (2016). Research Information Service Platform Design for University Libraries Based on User Needs. J. Modern Intelligence, 36 (04), 101-104+114.

[2] Grootendorst M. (2022). BER Topic: Neural topic modeling with a class-based TF-IDF procedure. J. arXiv preprint arXiv:2203.05794.

[3] Hu J G. (2011). A needs-based analysis of university library users. J. Library, (04), 137-138.

[4] Jafari S M. (2019). Abdollahzade Z. Investigating the relationship between learning style and game type in the game-based learning environment. J. Education and Information Technologies, 24(5), 1-22.

[5] Li, F L. (2014), Hierarchical analysis of user information requirements based on cognitive target classification. J. Knowledge Management Forum, (03), 19-23.

[6] Peng L. (2016). Three clues for defining the concept of "new media". J. Journalism and Communication Research, 23 (03), 120-125.

[7] Peng, Z., Wei, S., Tian, J., Qi, Z., & Bo, X., 2016. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).* IEEE.

[8] Qu W J, Li L Q, & Hu Y. (2016). Research on academic information exchange models in university digital libraries communities. J. Library Science Research, (17), 59-64+5.

[9] Ren L. (2016). The development of academic journals in the context of media integration. J. Academic Exchange, (03), 216-219.

[10] Sun R Y. (2014). Research on the correlation and interaction between information services and user cognitive processes. J. Intelligence Magazine, 33 (4), 190-195.

[11] Wang L. (2018). Library Makerspace User Information Service Demand Analysis and Service Strategy [J]. Library Information Work, 62 (12), 39-45.

[12] Zhen, W., Zhang, J., Feng, J., & Zheng, C., 2014. Knowledge Graph Embedding by Translating on Hyperplanes. In *National Conference on Artificial Intelligence*. AAAI Press.

[13] Zhang Y H & Yuan Q J. (2014) Advances in research on user information behavior in China. J. National Library Journal, 23 (06), 91-98.