Postprint: Multi-dimensional Feature Fusion for Academic Literature Download Behavior Prediction

Authors: Xie Hao, Wu Xuehua, Chen Qian, Tang Jing, white cloud, Mao Jin

Date: 2023-04-01T00:00:00+00:00

Abstract

[Purpose/Significance] Academic literature download behavior constitutes a crucial component of researchers' information retrieval activities. Research on predicting such behavior facilitates a deeper understanding of researchers' retrieval patterns, providing a basis for academic resource retrieval platforms to optimize search results and restructure ranking mechanisms, thereby enhancing the service quality of retrieval systems. [Method/Process] We construct a multi-dimensional feature framework for users' academic literature download behavior, develop sub-classifiers based on query relevance and user behavior using machine learning algorithms, and build a hybrid prediction model for academic literature download behavior through a weighted strategy. [Results/Conclusion] Experimental results demonstrate that the Random Forest algorithm achieves optimal performance on both classifiers. Compared to models trained solely on query relevance features, the hybrid model improves accuracy by 2.3% and F1-score by 1.3%. In the hybrid model, the user behavior-based sub-classifier carries higher weight; the features of "download volume", "whether professional/advanced search is employed", and "publication date" exhibit substantial contribution.

Full Text

Preamble

Volume 65, Issue 12, June 2021

Predicting Academic Literature Download Behavior by Integrating Multi-dimensional Features

Xie Hao, Wu Xuehua, Chen Xi, Tang Jing, Bai Yun, Mao Jin

Center for Studies of Information Resources, Wuhan University, Wuhan 430072

Abstract: [Purpose/Significance] Academic literature downloading represents a crucial component of researchers' information retrieval behavior. Predicting download behavior facilitates deeper understanding of researchers' retrieval patterns, providing a basis for academic resource platforms to optimize search results and restructure rankings, thereby enhancing retrieval system service quality. [Method/Process] This study constructs a multi-dimensional feature system for user academic literature download behavior, develops sub-classifiers based on query relevance and user behavior using machine learning algorithms, and builds a hybrid prediction model through weighted integration. [Result/Conclusion] Experimental results demonstrate that the Random Forest algorithm achieves optimal performance for both sub-classifiers. Compared with models trained solely on query relevance features, the hybrid model improves accuracy by 2.3% and F1-score by 1.3%. In the hybrid model, the user behavior-based sub-classifier carries higher weight, while "download volume," "use of professional/advanced search," and "publication time" emerge as highly contributive features.

Keywords: literature download prediction; multi-dimensional features; machine learning; hybrid model

Classification Number: G203

DOI: 10.13266/j.issn.0252-3116.2021.12.011

Academic resource retrieval platforms serve as critical channels for researchers to access scholarly information, offering advantages such as rich resources, timely updates, and convenient access. However, the rapid growth of academic resources has created information overload, increasing retrieval costs and consuming substantial time and effort from researchers. Optimizing retrieval functions through result restructuring and ranking is essential for improving user satisfaction with academic resource platforms and meeting researchers' scholarly information needs.

Academic literature downloading constitutes the subsequent process of academic retrieval. Understanding download behavior can inform academic search result ranking. Current research primarily focuses on the correlation between citation counts and download volumes [1-3], often treating download volume as a bibliometric evaluation metric to address the time-lag limitations of citation analysis. Some scholars have analyzed excessive downloading behavior from intellectual property perspectives, examining its characteristics [4-5], detection methods [6], and proposing countermeasures [7]. However, few studies predict academic literature download behavior, and existing work remains limited to forecasting download volumes [8], neglecting user preferences reflected in their interactive behaviors during academic retrieval [9] and failing to incorporate fine-grained individual search information into download predictions. Retrieval information can substantially reflect users' information needs, search objectives, and motivations [10]. Therefore, investigating influencing factors of academic literature download behavior within single retrieval sessions is significant for clarifying user

intent, optimizing academic literature retrieval result ranking, and improving researchers' retrieval efficiency and academic resource utilization.

Based on these considerations, we propose a multi-dimensional feature-based prediction model for academic literature download behavior. After constructing a relevant feature system for user download behavior, we employ machine learning algorithms to establish sub-classification models based on query relevance and user behavior, respectively, and construct a hybrid classification model using weighted strategies to predict users' academic literature download behavior.

1 Related Research

Academic search enables filtering and screening of scholarly information to meet researchers'diverse needs and personalized interests. Current scholarship primarily examines academic search behavior from perspectives of query intent, query formulation features, and retrieval strategies. Query intent refers to users' potential objectives during retrieval, which can be categorized as informational, navigational, or transactional [11]. Retrieval based on different query intents can reflect personalized differences among users [12]. M. Khabsa et al. [13] classify academic users' query intents into navigational and informational categories based on academic search behavior characteristics, where navigational intent targets specific academic literature and informational intent seeks relevant information on a topic [14]. As fundamental features of queries, query formulation construction reveals users' most direct needs [15]. X. Li et al. [16] found through query content analysis that academic searches predominantly involve entity retrieval, where these entities reflect topics of user interest. According to human information behavior theory [17], academic search behavior can be summarized into three types: research exploratory, task-oriented, and technique-dependent. Different academic search behavior types correspond to different retrieval strategies [18]. When users seek to understand domain development trends, they require large volumes of literature, exhibiting research exploratory behavior where keyword search and journal search are preferred strategies. When users have clear academic retrieval objectives, their behavior becomes task-oriented, tending toward exact matching modes.

Academic literature downloading represents the subsequent stage of academic retrieval. Researchers search for literature to locate scholarly information relevant to their work and download documents that meet their expectations. Studying academic literature downloads can compensate for the limitation that citation frequency cannot reflect the academic value of implicitly cited literature [19], considering contributions from read-but-uncited documents. Additionally, since papers undergo publisher review and reader comprehension before being cited, citation analysis suffers from time delays [20]. Download studies can mitigate this lag [21], more rapidly reflecting paper value [22]. Compared with citation data, download volume demonstrates greater discriminative power and sensitivity, with different statistical characteristics from citations [23]. Therefore,

download research can supplement citation behavior analysis, providing new perspectives for studying academic influence.

As an indicator of literature usage, historical download volume timely reflects usage patterns and can identify citation value earlier than citations [24]. Although the correlation between download frequency and citation frequency for individual open-access papers is not significant [25], the number of times papers from the *International Journal of Accounting Information Systems* enter the top 25 download rankings significantly correlates with citation counts [26], suggesting that the relationship between downloads and citations differs at the individual paper versus journal level. Furthermore, excessive downloading has become a common issue in university libraries. Xu Wenxian et al. [27] investigated excessive downloading cases domestically and internationally, finding that academic research needs and commercial interests constitute the primary causes.

In summary, existing research mostly evaluates academic influence by analyzing the relationship between download volume and citation count, or analyzes excessive downloading to regulate platform usage, lacking investigation into decision-making processes underlying user download behavior in academic retrieval. Therefore, we construct a user literature download behavior prediction model in academic retrieval from both query relevance and user behavior perspectives.

2 Hybrid Model for Academic Literature Download Behavior Prediction

2.1 Problem Definition

We define the prediction of literature download behavior within a single academic retrieval session as a binary classification problem: given user u and literature retrieval results $D = \{d_1, d_2, d_3, \cdots, d\}$ (where n is the total number of retrieved documents), for any document d in D, predict whether user u will download it. The prediction label y $\{0, 1\}$, where 1 represents download and 0 represents no download.

2.2 Academic Literature Download Behavior Prediction Model Framework

Academic literature download behavior is essentially driven by user needs, based on the fundamental assumption that literature more closely matching user needs is more likely to be downloaded. Based on previous research, we posit that this matching degree manifests in two aspects: (1) semantic similarity between the query and academic literature, and (2) similarity relationships between user needs reflected in behavior and literature. Accordingly, we first construct subclassifiers based on query relevance and user behavior for these two information types, then integrate them into a hybrid classifier. The query relevance-based

sub-classifier primarily learns document features and user query features, aiming to predict download behavior through matching degree between literature and user retrieval needs. The user behavior-based sub-classifier extracts literature embedding representations using the item2vec model [28] from user behavior records to mine potential associations between documents. The hybrid classifier applies weighted combination of predictions from these two sub-classifiers to comprehensively capture influencing factors and improve model effectiveness.

2.2.1 Query Relevance-Based Sub-Classifier Query relevance reflects the matching degree between retrieval results and user needs, directly influencing subsequent behaviors such as browsing and downloading, and represents an important basis for retrieval result ranking [29-30]. Query relevance assessment involves both literature and user query intent features.

Literature Features measure whether documents can satisfy user needs from quality and content perspectives. Information quality significantly influences users' perceived usefulness and subsequent behavioral decisions [31]. In academic information seeking contexts, high-quality academic literature enhances perceived usefulness, prompting download behavior. Commonly used literature quality metrics include citation count, download volume, source journal, and publication time [2, 32]. Citation count and download volume measure literature influence, while source journal and publication time reflect reliability and timeliness, respectively. Additionally, the degree to which literature content meets user information needs affects download decisions, measurable through matching degree between content and current query [33].

Query Intent reflects user search objectives and motivations, influencing subsequent browsing and downloading choices [34]. In academic information seeking, user goals may involve obtaining specific literature (navigational intent) or understanding publication volumes on a topic, institution, or author (informational intent), with the former more likely to result in downloads. Since query intent is abstract and difficult to identify directly, some studies indirectly reflect it through query formulation features [13, 34-35]. Following existing literature, we select five query features: query length, whether it is a title, search fields, exact matching usage, and professional/advanced search usage.

Feature extraction methods are as follows:

(1) Literature Feature Extraction. Literature metadata features can be directly obtained from bibliographic information. The source journal feature identifies whether the literature comes from authoritative journals (e.g., those indexed by Peking University Core, CSSCI, CSTPCD, EI, SCI). Publication time is calculated as the difference between publication year and current browsing/download year. Due to the highly skewed and uneven distribution of citation counts and download volumes, we apply equal-frequency binning to process them. For citation counts, three bins are created: low [0, 1], medium (1, 5], and high (5, $+\infty$). Download volume is similarly binned into low [0, 27], medium

(27, 95], and high (95, $+\infty$).

For literature matching degree calculation, since a query may contain multiple search fields, we divide the process into two parts: (1) For content-related fields (e.g., topic, title), we use keyword matching to calculate the matching degree between search terms and literature content. Specifically, we segment the query, literature title, and abstract, then calculate the proportion of search terms appearing in the title, abstract, or keywords. For example, if a query contains three terms and two appear in the title, keywords, or abstract, the matching degree is 0.66. (2) For metadata-related fields (e.g., author, journal), if the query contains fields that exactly match the literature metadata, we add a value between 0 and 1 to the content matching degree as a hyperparameter tuned during training. The formula is:

$$score_{total} = \begin{cases} score_{content} & \text{if no matching metadata fields exist} \\ score_{content} + \beta(\beta \in [0,1]) & \text{if matching metadata fields exist} \end{cases}$$

where $score_{total}$ is the overall matching degree, $score_{content}$ is the content matching degree, and β is the added value.

(2) User Query Feature Extraction. Query length is obtained by segmenting and counting terms. Users submitting longer queries typically seek more specific information. Title identification follows this rule: if the number of matching words between the query and article title exceeds 5, the query is considered a title; otherwise, it is not. Direct title searching indicates users likely seek specific literature. The search field feature checks whether the query contains author, DOI, or title fields, which suggest clear search intent. Exact matching, advanced search, and professional search are identified through regular expressions. Queries with exact matching typically contain quotation marks, while advanced/professional searches usually include logical operators (“and,”“or,”“not,”“*,”“+,”“^”) or order operators (“(”). Different search strategies can reflect user intent, preferences, and information need types to some extent.

Table 1 summarizes the feature system for the query relevance-based subclassifier.

2.2.2 User Behavior-Based Sub-Classifier Recommendation research indicates that user browsing or purchase sequences contain product similarity information. Product embedding representations trained from user behavior sequences can map original high-dimensional sparse data into low-dimensional feature spaces, making similar products close in spatial distance, thereby modeling potential relationships and improving recommendation effectiveness [36-38]. We adopt this approach, training low-dimensional literature embedding representations from user-literature interaction records to capture deep associations between documents.

A key approach for training product embeddings in recommendation involves adapting the word2vec model from natural language processing [28]. Specifically, products are treated as words in word2vec, user browsing/purchase sets as sentences, and product pairs appearing in the same set as positive samples. The Skip-gram with Negative Sampling (SGNS) model learns low-dimensional product embeddings (item2vec). We treat literature as words and each user' s browsed document set within a timeframe as a sentence. Formally, using *article_{id}* to identify a document, a user' s literature browsing records can be represented as:

$$user_i = [article\_id1, article\_id2, ..., article\_idm]$$

The training data for the item2vec model is represented as:

$$train\_list = [user_1, user_2, ..., user_i, ..., user_n]$$

where n represents the number of users.

We input these user browsing literature sets into the SGNS model to learn low-dimensional literature embeddings:

$$item\_embedding = SGNS(train\_list)$$

The resulting literature embeddings are then input into a classifier to output predictions from the sub-classifier.

2.2.3 Hybrid Classification Model The hybrid classification model weights the download/non-download probabilities predicted by the two sub-classifiers to obtain final predictions. Weight coefficients are determined during model tuning.

In summary, the proposed academic literature download behavior prediction framework, illustrated in Figure 1, comprises three classifiers. For input user search and browsing records, the model: (1) builds query relevance-based and user behavior-based sub-classifiers using machine learning algorithms; (2) constructs a hybrid classifier from the two sub-classifiers; and (3) predicts whether a user will download a specific document based on the hybrid classifier' s output.

3 Experiments and Results Analysis

3.1 Experimental Design

The experimental process, shown in Figure 2, consists of: (1) **Data field expansion**: A crawler was developed to collect literature and journal information, expanding source data fields. (2) **Data preprocessing**: The expanded dataset underwent association and invalid data removal, producing usable experimental

data split into training and test sets at an 8:2 ratio. (3) **Feature extraction**: Literature features and user query features were extracted from training data using the method in Section 2.2.1. Simultaneously, the item2vec model was trained on non-institutional user browsing data to extract literature embeddings based on user behavior records. (4) **Model training**: The training data was used to train both sub-classifiers. Mature machine learning algorithms for classification include Logistic Regression (LR), Decision Tree (DT), Naive Bayes (NB), Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), and Random Forest (RF). We experimented with these algorithms, selecting the best-performing one for the final hybrid classifier. (5) **Model prediction and fusion**: Test set features were extracted similarly. Literature and query features were input to the query relevance-based sub-classifier to obtain prediction p_p^i, while literature embeddings were input to the user behavior-based sub-classifier to obtain prediction p_u^i. These predictions were weighted to build the hybrid classifier:

$$p_i = \alpha \cdot p_p^i + (1 - \alpha) \cdot p_u^i$$

where α is the model fusion weight coefficient, i represents non-institutional user behavior records when p_u^i is from non-institutional data, and institutional user behavior records otherwise.

3.2 Dataset and Preprocessing

Our research data comes from the Wanfang Data Knowledge Service Platform journal literature user behavior log dataset provided by the "Huiyuan Sharing" National University Open Data Innovation Research Competition, including 37,544,670 search logs, 11,998,421 browsing logs, and 14,025,159 download logs spanning December 1, 2019 to January 31, 2020 [39]. Since source data fields were insufficient, we expanded user browsing logs by crawling additional information using the title and author fields for advanced searches on the Wanfang platform, capturing top-ranked paper information. Expanded fields included literature metadata (abstract, publication year, download count, citation count) and journal metadata (source journal ID, journal level, name, total downloads, total citations, impact factor).

A challenge with the source dataset is that search, browse, and download records are stored separately, and institutional users cannot be individually identified. Analyzing only non-institutional users would waste substantial data resources. To incorporate institutional user records, we concatenated logs using: (1) minimum time difference between browsing and search times for the same user (*user_{id}*) and same document (*article_{id}*); (2) keyword co-occurrence methods to assess relevance between search terms and document titles after segmentation and stopword removal; (3) minimum time difference between download and browse times for the same user and document. After removing records

with missing publication years or abstracts, we obtained 2,383,933 experimental records (Table 2).

Users typically exhibit two download behavior patterns (Figure 3): (1) browse detailed information (abstract, etc.) before deciding to download (search-browse-download/not download); (2) directly judge from brief retrieval list information whether literature meets needs (search-download/not download). Our experimental data comprises post-search browsing records, with positive samples being search-browse-download records and negative samples being search-browse-not download records.

3.3 Evaluation Metrics

We selected accuracy, recall, precision, and F1-score as evaluation metrics:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1 = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

where TP is true positives (predicted and actual downloads), TN is true negatives (predicted and actual non-downloads), FP is false positives (predicted download but actual non-download), and FN is false negatives (predicted non-download but actual download).

3.4 Experimental Results Analysis

Tables 3, 4, and 5 present prediction results for the query relevance-based sub-classifier, user behavior-based sub-classifier, and hybrid classifier, respectively.

Results show Random Forest achieves optimal performance across all metrics for both sub-classifiers. Notably, in the user behavior-based sub-classifier, Random Forest' s precision exceeds the second-best Decision Tree by 14 percentage points, reaching 78.3%, but with low recall. The query relevance-based sub-classifier shows the opposite pattern. Fusion creates complementary effects, improving overall performance. Table 5 confirms the hybrid classifier outperforms sub-classifiers, with accuracy increasing 2.3% and F1-score increasing 1.3% compared to the query relevance-based model alone.

3.5 Sub-Classifier Weight Analysis

To investigate weight distribution between sub-classifiers in the hybrid model, we used exhaustive search to explore optimal weight coefficient α (range [0, 1], step 0.1). Results in Figure 4 show accuracy peaks at $\alpha = 0.3$, while F1-score peaks at $\alpha = 0.4$. Considering both metrics, the optimal α is 0.4, where accuracy is near its peak and F1-score is optimal. This indicates the user behavior-based sub-classifier carries higher weight than the query relevance-based sub-classifier in the optimal hybrid model, likely because it leverages historical behavior records to extract fine-grained intrinsic features and learn potential literature relationships.

3.6 Feature Contribution Analysis

Feature contribution analysis reveals discriminative capabilities and enhances model interpretability. For query relevance features, we calculated information gain-based contributions (Figure 5). Features with contribution >10% are considered high-impact, including "download volume,""use of professional/advanced search," and "publication time."

Download volume significantly impacts classification. As shown in Figure 6(a), small download volumes correlate with higher non-download rates after browsing, while large download volumes correlate with higher download rates, reflecting researchers' tendency to judge quality by download count—consistent with the Matthew effect in information resource distribution.

Use of professional/advanced search and **query length** reflect search goal clarity. When users employ professional/advanced search or submit long queries, their search purpose is explicit and requirements are stringent, reducing download likelihood. Conversely, short queries suggest interest in domain overview, requiring large literature volumes and increasing download probability (Figures 6(b) and 6(e)).

Publication time reflects content novelty. Recent publications show higher post-browse download rates, while older publications show higher non-download rates (Figure 6(c)), indicating researchers prefer recently published articles.

Literature matching degree reflects content-user need alignment. Higher matching correlates with higher download rates, while lower matching correlates with non-downloads (Figure 6(d)). We explored optimal β values for metadata field contributions via exhaustive search but found minimal performance impact, likely due to few records containing metadata-related search fields, making content-based matching dominant.

Low-contribution features include **source journal** and **citation count**, suggesting users pay limited attention to journal authority and citation magnitude when making download decisions. Research also indicates no significant correlation between download and citation volumes at individual paper level [40].

Search fields, **exact matching usage**, and **title query** show low contribution with no significant positive/negative sample distribution differences across value ranges (Figures 6(g), 6(h), 6(j)), possibly because these query formulation features inadequately distinguish query intent types, warranting future exploration of more precise query intent identification.

Conclusion

We constructed a multi-dimensional feature system for academic literature download behavior and built query relevance-based and user behavior-based prediction sub-models using machine learning, weighted for fusion. Experiments on Wanfang user behavior logs demonstrate Random Forest's superior performance and the hybrid classifier' s effectiveness. The user behavior-based sub-classifier carries higher weight in the hybrid model. "Download volume," "use of professional/advanced search," and "publication time" emerge as key influencing factors. The proposed model achieves good performance and can serve as a re-ranking module in retrieval systems: after returning relevant literature sets based on user queries, the model extracts features from user behavior and literature data, predicts downloadable documents through trained classifiers, and ranks these at the top of result pages to enhance retrieval efficiency.

Limitations include: (1) using query features only indirectly reflects query intent without building explicit intent recognition models—future work should integrate intent recognition with download prediction; (2) dataset limitations prevent individual identification within institutional users, precluding session analysis. Future research will incorporate session sequences to analyze temporal evolution of content and behavior preferences during academic retrieval, improving and innovating the prediction model.

References

[1] Xiong Zequan, Duan Yufeng. Can early download volume predict later citation count?—Taking library and information science journals as examples[J]. Library and Information Knowledge, 2018(4): 32-.

[2] Xie Juan, Gong Kaile, Cheng Ying, et al. Meta-analysis of the correlation between paper download volume and citation count[J]. Journal of the China Society for Scientific and Technical Information, 2017, 36(12): 1255-1269.

[3] Wang Chao. Research on the relationship between journal paper citation volume and download volume[J]. Information Research, 2020(6): 33-39.

[4] Zhang Wei, Dai Guoqiang. Analysis of characteristics and countermeasures for excessive database downloading in domestic university libraries[J]. Office Automation, 2016, 21(14): 25-27.

[5] Zhang Min, Zhang Lei. Enabling and inhibiting factors of excessive digital library e-resource downloading intention[J]. Library Science Research, 2016(16):

51-57, .

[6] Zhang Min, Zhang Lei. Research on influencing factors of excessive digital library e-resource downloading intention—Based on dual contexts of task-driven and punishment inhibition[J]. Library and Information Service, 2016, 60(7): 116-122.

[7] Sun Li. Research on university students' database paper downloading behavior in Guangzhou University Town[J]. Library and Information Science Journal, 2016, 1(11): 150-153.

[8] Liu Ying. Predicting paper download volume based on ARIMA model and neural network[D]. Dalian: Dalian University of Technology, 2015.

[9] LI X, DERIJKEM. Characterizing and predicting downloads in academic search[J]. Information processing & management, 2019, 56(3): 394-407.

[10] Zhang Haitao, Zhang Xiaohui, Wei Ping, et al. Research progress on network user information retrieval behavior[J]. Information Science, 2020, 38(5): 169-176.

[11] BRODER A. A taxonomy of Web search[J]. SIGIR forum, 2002, 36(2): 3-10.

[12] DOU Z, SONG R, WEN J. A large-scale evaluation and analysis of personalized search strategies[C]//Proceedings of the 16th international conference on World Wide Web. New York: ACM, 2007: 581-590.

[13] KHABSA M, WU Z, GILES C L. Towards better understanding of academic search[C]//Proceedings of the 16th ACM/IEEE-CS on joint conference on digital libraries. New York: ACM, 2016: 111-114.

[14] Zhang Xiaojuan. Comparative analysis of personalization potential for informational, navigational, and transactional queries[J]. Digital Library Forum, 2017(9): 35-41.

[15] Wu Dan, Sun Haodong. Comparative analysis of mobile library WAP and APP user retrieval behavior[J]. Library and Information Service, 2016, 60(18): 14-20.

[16] LI X, SCHIJVENAARS B J A, DERIJKEM. Investigating queries and search failures in academic search[J]. Information processing & management, 2017, 53(3): 666-683.

[17] WILSON T D. Human information behavior[J]. Informing science, 2000, 3(2): 49-56.

[18] Wang Jiandong, Wang Jimin. Research on university user journal database retrieval behavior based on log mining[J]. Information Science, 2009, 27(5): 690-694.

[19] Lou Haiping, Pan Xingmei, Fang Hong, et al. Review of Chinese academic paper download metrics research[J]. Library Research and Work, 2018(10): 50-55.

[20] Guo Qiang, Zhao Jin, Liu Siyuan, et al. Research on statistical properties of scientific paper download counts[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2012, 48(1): 29-.

[21] GARFIELD E. Fortnightly review: How can impact factors be improved?[J]. BMJ, 1996, 313(7054): 411-413.

[22] Zhao Yiquan, Wang Zhenmin, Xiong Wenbing, et al. Research on the relationship between scientific paper downloads and citations—Taking ACM Digital Library as an example[J]. Chinese Journal of Scientific and Technical Periodicals, 2014, 25(6): 818-823.

[23] Zhao Xing. Research on measurement characteristics of academic literature usage data[J]. Journal of Library Science in China, 2017, 43(7): 44-57.

[24] Yang Li, Xiong Zequan, Duan Yufeng. Research on journal paper citation prediction based on quantile regression[J]. Information Science, 2019, 37(10): 60-66.

[25] Niu Yuxin, Zong Qianjin, Yuan Qinjian. Bibliometric study of open access paper downloads and citations[J]. Journal of Library Science in China, 2012, 38(4): 119-127.

[26] O' LEARY D. On the relationship between citations and appearances on "top 25" download lists in the international journal of accounting information systems[J]. International journal of accounting information systems, 2008, 9(1): 61-75.

[27] Xu Wenxian, Chen Xuemei. Research on excessive database downloading behavior in university libraries[J]. Library Theory and Practice, 2014(11): 20-23.

[28] BARKAN O, KOENIGSTEIN N. Item2vec: neural item embedding for collaborative filtering[C]//2016 IEEE 26th international workshop on machine learning for signal processing. Piscataway: IEEE, 2016: 1-6.

[29] Yang Shuxin, Xu Huiqin, Tan Wei. Keyword query ranking method combining query relevance[J]. Computer Engineering and Design, 2013, 34(9): 3136-3140.

[30] Wu Lihua, Luo Yunfeng, Zhang Hongbin. Research on information retrieval models and relevance algorithms[J]. Journal of Intelligence, 2006(12): 25-27.

[31] Zhang Liyi, Zhang Ran. Analysis of antecedents of key variables in Technology Acceptance Model (TAM)[J]. Journal of Information Resources Management, 2015, 5(2): 11-20.

[32] Wang Haitao, Tan Zongying, Chen Ting. Research on factors influencing paper citation frequency—On the rationality of using citation frequency to evaluate research quality[J]. Studies in Science of Science, 2016, 34(2): 171-177.

[33] Shen Min, Yang Xinyao, Wang Kai. Research on university library user preference retrieval system based on machine learning[J]. Library and Information Service, 2015, 59(11): 143-148.

[34] Lu Wei, Zhou Hongxia, Zhang Xiaojuan. Review of query intent research[J]. Journal of Library Science in China, 2013, 39(1): 100-111.

[35] BELKIN N J, KELLY D, KIM G, et al. Query length in interactive information retrieval[C]//Proceedings of the 26th annual ACM SIGIR conference on research and development in information retrieval. New York: ACM, 2003: 205-212.

[36] TANG J, WANG K. Personalized Top-N sequential recommendation via convolutional sequence embedding[C]//The eleventh ACM international conference. New York: ACM, 2018.

[37] 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.

[38] ZHANG W, DU Y, YOSHIDA T, et al. Deep Rec: a deep neural network approach to recommendation with item embedding and weighted loss function[J]. Information sciences, 2019, 470(2019): 121-140.

[39] "Huiyuan Sharing" National University Open Data Innovation Research Competition Organizing Committee. "Huiyuan Sharing" National University Open Data Innovation Research Competition—Submission guidelines for entries[EB/OL]. [2020-07-01]. http://hdl.handle.net/20.500.12291/10232V2[Version].

[40] Lu Wei, Qian Kun, Tang Xiangbin. Correlation between literature download frequency and citation frequency—Taking library and information science as an example[J]. Information Science, 2016, 34(1): 3-.

Author Contributions: Xie Hao: model construction, experimental code writing, paper drafting and revision; Wu Xuehua: feature system construction, model construction, paper drafting and revision; Chen Xi: feature system construction, model construction, paper drafting and revision; Tang Jing: model construction, paper drafting and revision; Bai Yun: model construction, paper drafting and revision; Mao Jin: proposed research ideas and paper revision suggestions.

Received: 2020-11-26 **Revised:** 2021-03-03 **Pages:** 112-121 **Responsible Editor:** Xu Jian

Citation: Xie Hao, Wu Xuehua, Chen Xi, et al. Predicting download behavior of academic literature based on multi-dimensional features[J]. Library and Information Service, 2021, 65(12): 112-121.

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