Precision Recommendation of Knowledge Discovery Services Based on User Interest Metrics (Postprint)

Authors: Ding Mengxiao, Bi Qiang, Xu Pengcheng, Li Jie, Mou Dongmei

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

Abstract

[Purpose/Significance] To address the issues of low personalization and suboptimal recommendation effectiveness prevalent in current knowledge discovery services, this paper proposes a recommendation algorithm grounded in user interest measurement and content analysis. [Method/Process] An academic resource model is constructed through three dimensions: feature word distribution, LDA topic distribution, and citation structure network. By quantifying user behavior, the algorithm computes users' interest degree in browsed academic resources, and synthesizes this with the academic resource model to build a user interest model. The user interest model is then matched against the academic resource model to calculate similarity, yielding users' interest values for each academic resource. Finally, the TOP-N academic resources with the highest interest values are recommended to users. [Results/Conclusion] Experimental validation demonstrates the algorithm's effectiveness and recommendation accuracy. The results indicate that, from the perspective of real-time dynamic interest measurement, the proposed recommendation algorithm can effectively predict user interests, achieves significant recommendation performance, and offers insights for realizing precise recommendations in discovery services.

Full Text

Precise Recommendation of Knowledge Discovery Services Based on User Interest Measurement

Ding Mengxiao[1], Bi Qiang[1], Xu Pengcheng[1], Li Jie[1], Mu Dongmei[2]
[1]School of Management, Jilin University, Changchun 130022 [2]School of Public Health, Jilin University, Changchun 130021

Abstract

[**Purpose/Significance**] To address the current issues of low personalization and poor recommendation effectiveness in knowledge discovery services, this paper proposes a recommendation algorithm based on user interest measurement and content analysis. [**Method/Process**] The paper constructs an academic resource model through three dimensions: feature word distribution, LDA topic distribution, and citation structure network. By measuring user behavior, it calculates users' interest in browsed academic resources and combines this with the academic resource model to build a user interest model. The user interest model is then matched with the academic resource model to calculate similarity, obtaining interest values for each academic resource, and finally recommending the TOP-N academic resources with highest interest values to users. [**Result/Conclusion**] Experimental verification demonstrates the algorithm's effectiveness and recommendation accuracy. Results show that from the perspective of real-time dynamic interest measurement, the proposed recommendation algorithm can better predict user interests with significant recommendation effects, providing insights for achieving precise recommendation in discovery services.

Keywords: user interest; content analysis; discovery service; precise recommendation
Classification Number: G251
DOI: 10.13266/j.issn.0252-3116.2019.03.003

We have transitioned from the digital era to a data-driven era where data serves as both an asset and a resource. Faced with exponentially growing, richly typed massive data resources, effective utilization to achieve user-oriented knowledge service innovation has become a key research focus. Knowledge discovery services, as a crucial component of knowledge services, serve as the vital link connecting resources and users. The breakthrough point for enhancing knowledge discovery service capabilities lies in accurately grasping user interest preferences, predicting user needs, discovering required knowledge, and proactively recommending it to users. However, current knowledge discovery systems suffer from insufficient flexibility, inaccurate recommendations, and low levels of personalized service. Meanwhile, a fusion gap exists between users' usage environments and knowledge service environments, and active resource discovery services integrated into user environments need strengthening. In the data-driven era, "precision service" represents the development direction across industries, aligning with the development 思路 of digital library knowledge discovery services. Precise recommendation is an important means to improve knowledge discovery service quality. Integrating precise recommendation models into knowledge discovery systems, analyzing user behavior to determine interests, aggregating related resources, and fully utilizing new technological means to provide knowledge services for users is essential for leveraging discovery systems' advantages

in data resources and meeting personalized, precise, and knowledge-based user needs. This paper analyzes user behavior sets in discovery systems to identify dynamic interests, constructs user interest models and resource content models, and provides precise knowledge recommendations for users through relevant algorithms, thereby using precise recommendation technology to change the interaction mode between users and knowledge discovery systems and provide accurate recommendation services.

1 Related Research

Since the emergence of the web resource discovery system Summon in 2009, knowledge discovery services based on discovery systems have developed for nearly a decade. Research during this period has primarily focused on discovery service concepts, functional analysis, system comparisons, and applications. Studies at both theoretical and functional application levels reflect academic and industry expectations for improving knowledge discovery service quality, though research from a precise recommendation perspective remains limited.

As personalization becomes the trend in knowledge services, knowledge discovery systems must provide precise recommendations to enhance service quality. Precise recommendation is the critical step in discovery services and users' deepest experience of the services provided by discovery systems. Research on recommendation services both domestically and internationally focuses on the importance of precise recommendation and recommendation algorithms. Google Scholar, launched in 2004 as a metadata repository-based discovery service, supports related article recommendations and has generated enthusiastic responses. Scholars have subsequently invested in using personalized recommendations to improve knowledge service quality. S.Q. Yang and K. Wagner argued that discovery systems should have the service function of recommending related resources, implementing the concept of "service to people" to provide thematic push for users. When evaluating knowledge discovery systems EDS and Summon, Lynchburg College comprehensively considered various factors affecting discovery system performance and quality, listing similar search result recommendations ("more like this") as an important indicator. Qin Hong believed that digital library resource discovery systems should have characteristics such as personalized discovery and automated recommendation, providing users with precise knowledge push when facing interactive contexts. Zhang Jun constructed user portraits based on users' basic information and behavior information to predict user needs preferences and potential knowledge needs, thereby achieving personalized recommendation matching for knowledge discovery.

If knowledge discovery services discover new knowledge and hidden associations between knowledge for users, then precise push services provide users with a specialized acquisition and application pathway for utilizing knowledge. Through precise recommendation, a win-win situation for discovery system services and users can be achieved. Precise recommendation is an indispensable part of knowledge discovery systems and a way for users to obtain quality service expe-

riences.

Precise recommendation services require quality recommendation algorithms. Currently, mainstream recommendation algorithms for academic resources include content-based recommendation algorithms and collaborative filtering recommendation algorithms based on user rating matrices. Content-based recommendation algorithms extract content features of academic resources and recommend similar resources to users, with clear and intuitive results. P. Guan et al. used content recommendation methods to enhance semantic information of scientific literature by merging metadata such as titles, keywords, abstracts, and citations, using TF-IDF algorithms to obtain thematic word weight vectors to build user interest models and improve recommendation interpretability. However, F. Ricci et al. pointed out that content-based recommendation methods only consider resource content features, focusing on the meaningfulness, structure, and extractability of characteristic content, without fully considering user interests, thus not fully achieving personalization goals.

Collaborative filtering recommendation algorithms cluster users through behavior data and interests, with the basic principle of gathering users with the same interests. Users with similar user-item ratings are considered to have the same interests. When "neighbor" users have browsed an item that the target user has not, that item is recommended to the target user. Collaborative filtering strategies require user rating information and perform well in e-commerce and other fields with wide applications. However, the most challenging problems for collaborative filtering algorithms are sparsity and cold start—when a system has insufficient user rating data or information, recommendation effectiveness is greatly reduced. The question of how the first user discovers new items remains to be solved, and collaborative filtering algorithms cannot effectively provide precise recommendations for new users and new items, failing to meet personalized user needs. In the digital library field, where user initiative is poorer than in e-commerce, when user interaction drive with digital libraries is insufficient, the sparsity and cold start drawbacks of collaborative filtering algorithms are amplified.

With prominent issues such as data silos, information overload, and information disorientation, the chronic problems of traditional recommendation systems and algorithms remain unresolved, leading to reduced user satisfaction and even user loss, constraining further promotion and application of recommendation services. Traditional recommendation algorithms cannot keep pace with the speed of user interest and preference changes, making research on recommendation algorithms that dynamically capture users' potential interests particularly important. Therefore, information service providers must fully consider users' dynamic interests, meeting users' dynamic knowledge needs based on user interest measurement and content analysis to provide more precise knowledge discovery services. In view of this, this paper aims at precise recommendation, combining existing algorithms, leveraging the advantages of massive resource data and user data in knowledge discovery systems, and conducting research on

current cold start problems and user interest shift issues in recommendations, systematically describing users' implicit interests when browsing academic resources, identifying users' current interest states, and proposing an innovative recommendation algorithm based on user interest models.

2 Resource Content Measurement

The key to precise recommendation lies in accurately grasping user needs, interests, or preferences, deeply mining resource content features, establishing connections between users and resources, and providing personalized knowledge recommendation services. Therefore, in precise recommendation, we must first establish an academic resource model, then combine it with user interest values to build a user interest model, and finally determine how to use the user interest model for precise recommendation. Academic resource modeling primarily involves extracting text features, with feature words, thematic words, and citations being the main text features of academic resources. By extracting feature word distribution, thematic word distribution, and citation structure networks of academic resources, we construct the academic resource model. We define M_d as the academic resource model, K_d as the feature word distribution of academic resources, T_d as the thematic distribution of academic resources, and C_d as the citations of academic resources. The academic resource model is represented as $M_d = \{K_d, T_d, C_d\}$.

2.1 Feature Word Distribution

Define the document feature word set as $K = \{K_{d1}, K_{d2}, \ldots, K_{dn}\}$, where d represents an academic text. The common method for text feature word extraction is the TF-IDF algorithm, which calculates the TF-IDF value of words in documents. Words with larger TF-IDF values can serve as document feature words. However, the traditional TF-IDF algorithm cannot capture distribution proportion differences of words in text collections, which are important factors for expressing text content. Therefore, we introduce the concept of information gain to improve the traditional TF-IDF algorithm. The weight W_{di} of feature word K_{di} is calculated as shown in Formula (1):

Formula (1)

where TF_{di} represents the frequency of the i-th feature word in academic document d, N represents the total number of academic documents, and n_d represents the number of academic resources containing feature word i.

The IG_d in Formula (1) is information gain, representing the information quantity of words, calculated as shown in Formula (2):

Formula (2)

where $H(d) = \sum(P(i) \times \log_2 P(i))$, $H(d|i) = -\sum(P(d|i) \times \log_2(P(d|i)))$, $P(i) = |wf(i)|/\sum|wf|$. $|wf(i)|$ represents the sum of word frequencies in document d.

The vector $K_d = \{(K_{d1}, W_{d1}), (K_{d2}, W_{d2}), \ldots, (K_{dn}, W_{dn})\}$ is called the feature word distribution of academic resources.

2.2 Topic Distribution

Define the topic distribution as $T = \{T_{d1}, T_{d2}, \ldots, T_{dn}\}$, where d is an academic resource and T_{di} represents the topic distribution probability of document d_i. The topic distribution of academic resources uses the LDA algorithm to obtain the joint distribution probability of documents' topics and feature words $p(w|d) = p(w|t) \cdot p(t|d)$. Using Gibbs sampling to solve the posterior parameters of the LDA model, $P(T_{di}|d)$ represents the posterior probability that academic resource d belongs to topic T_{di}. The vector $T_{di} = \{P(T_{d1}|d), P(T_{d2}|d), \ldots, P(T_{dn}|d)\}$ is called the LDA topic distribution of academic resources.

2.3 Citation Structure

Define C_d as the citations of academic resources, with the citation set represented as $C_d = \{C_{d1}, C_{d2}, C_{d3}, \ldots, C_{dn}\}$. The mutual citation relationships among scientific literature implicitly contain similarity relationships between documents. Through citation relationships, a series of content-related documents can be found to serve recommendation systems. Citation relationships can be established based on scientific citation relationships, using graph theory to construct citation graphs. Generally modeled as graph $G = (V, E)$, the vertex set V is the information object set, where any point $d_i \in V$ represents a citation. The edge set E represents relationships between vertices. If citation d_i cites citation d_j, this relationship is represented by edge $(d_i, d_j) \in E$ (see Figure 1 [Figure 1: see original paper]). Using graph theory methods to mine implicit relationships between vertices in citation structure graphs, citation structure similarity is calculated using topological structure information of graphs.

3 User Interest Measurement

User interest preferences are the main basis for recommendation systems to recommend resources. The accuracy of user interest measurement directly affects the quality of precise recommendation in knowledge discovery services. Liu Hongwei et al. quantified users' dynamic implicit interests to help personalized recommendation services in e-commerce. Zeng Ziming et al. discriminated dynamic changes in user interests from a user experience perspective for knowledge recommendation services in digital libraries. Related research shows that accurate measurement of user interests can effectively improve recommendation quality in knowledge discovery services and produce more precise recommendation effects. Analyzing and describing user interests is the foundation for knowledge discovery services to achieve precise recommendation, which can be

realized by establishing user interest models. User interest models describe users' interest preferences for resource information, comprehensively reflecting users' demand levels for resource information within a certain period through analysis of user interests.

3.1 Behavior Measurement

User interests can be divided into explicit and implicit interests. Explicit interest refers to users actively providing their interest tendencies for knowledge needs, mainly derived from personal information filled in during registration. Implicit interest refers to interest preferences implied behind various user behaviors when using the system. Because explicit interest is usually stable and users have poor participation initiative, with characteristics of inaccuracy, incompleteness, and subjectivity, it cannot reflect users' dynamic interests. In contrast, implicit interest does not require users' explicit participation during data collection; data only needs to be recorded when users generate behaviors, without affecting user browsing. Therefore, this paper uses implicit interest to dynamically measure user interests. When users browse, the system automatically tracks and records user behavior data on the server side, calculates users' interest in page content based on behavior data, and obtains users' interested themes and content. Through mining user behavior data, the obtained user interests are more objective and accurate.

User implicit interest measurement is primarily based on user browsing behavior. R. Krishnamoorthy believed that user interest measurement is based on combinations of user browsing behaviors, dividing browsing behaviors into verification behaviors and activation behaviors. Verification behaviors can be used to determine whether users have interests, such as saving pages, printing pages, and visiting the same page multiple times. These behaviors show whether users are interested in browsing themes or pages, enabling user behavior data collection to determine interest levels. Activation behaviors are the next stage of verification behaviors, referring to behaviors that can determine user interest levels, such as browsing time on pages, mouse activities, and keyboard activities. L. Zheng et al. comprehensively analyzed the relationship between user browsing time and user interest, proposing a method to calculate thematic interest through user browsing time, confirming that the algorithm is reasonable and accurate for calculating user interest and can be effectively used for personalized recommendation. Browsing duration is an important behavioral manifestation of users' interest in browsing content; the longer users browse, the higher their interest in page content.

When users are interested in opened pages and their content or consider them valuable, they will spend longer time browsing. If users are not interested in browsing content, they will quickly close pages and click next pages to search for interesting content. Main factors affecting user browsing time include: (1) User attention to content. Higher attention to content or themes leads to longer browsing time. (2) Page content volume. Larger information capacity on pages

may lead to longer time spent. (3) User comprehension ability. When two users have the same attention level to the same page content, stronger comprehension ability leads to shorter browsing time. Therefore, user interest in pages should be measured through vertical comparison of the same user. Due to individual differences among users, using absolute browsing time as a measurement basis for user interest in a particular page is biased. Instead, the ratio of relative time spent by the same user browsing different pages, combined with the absolute ratio of information volume on different pages, should be used as the benchmark for measuring user interest.

Additionally, considering user interest transfer and to reflect recent learning progress and interests, we select academic documents browsed by users within a period T and other interaction behaviors for measurement. Besides browsing duration and page information volume, if users are particularly interested in an academic resource or consider it particularly valuable, they will further generate interaction behaviors such as downloading, collecting, and sharing. These interaction behaviors better reflect users' implicit interests and should have higher weights in interest measurement. Based on these considerations, user interest in academic documents is calculated as shown in Formula (4):

Formula (4)

where $UI_i|time|$ represents users' effective browsing time for academic document i, and $D_{icontent}$ represents the content volume of academic document i, which can be expressed by the byte size of the academic document. T_{min} is a small value designed to prevent mis-clicks. If users' browsing time for academic document i is less than T_{min}, it is considered a mis-click and $UI_{iInterest} = 0$. If greater than or equal to T_{min}, user interest in document i is calculated through Formula (1). If users have interaction behaviors such as downloading, collecting, or sharing with academic document i, user interest increases, and $UI_{iInterest}$ calculated through Formula (1) increases by δ, where δ is a tuning parameter set to 1 in this paper.

3.2 User Interest Model

Based on the academic resource model, we construct the user interest model. Define K_u as the feature word vector of user interest, T_u as the user's preferred topic distribution, and C_u as the citation distribution of browsed literature. The user interest model can be represented as $M_u = \{K_u, T_u, C_u\}$.

3.2.1 Feature Word Preference An academic resource in a knowledge discovery system often contains multiple feature words that can briefly summarize and describe the resource content. Let $\{d_1, d_2, d_3, \dots, d_i\}$ represent the set of all academic resources browsed by a user within period T. Using segmentation tools and corpora, we extract the feature word set $K_d = \{K_{d1}, K_{d2}, \dots, K_{dn}\}$ of academic documents browsed by the

user. The user's feature word preference can be described by a vector $K_u = \{(K_{d1}, W_{d1}), (K_{d2}, W_{d2}), (K_{d3}, W_{d3}), \ldots, (K_{di}, W_{di})\}$, where K_{di} represents the i-th preference feature word and W_{di} is the weight of feature word K_{di}. The weight W_{di} of text feature words directly uses the calculation results from the TF-IDF+IG algorithm in the academic resource modeling above. Then:

$$K'_{ui} = UI_{iInterest} \cdot K_{ui}$$

where K_{ui} is the feature word distribution of academic resources, $UI_{iInterest}$ represents the user's interest in the i-th feature word, and K'_{ui} represents the new feature word vector of academic resources browsed by the user.

3.2.2 Topic Preference The set of academic resources of a certain type browsed by a user within period T is $\{d_1, d_2, d_3, \ldots, d_i\}$. The user's LDA topic preference can be described by an N-dimensional vector $T_u = (T_{u1}, T_{u2}, T_{u3}, \ldots, T_{un})$.

$$UI_{iInterest} \times T_{di}$$

where T_{di} is the topic probability distribution of academic resources, $UI_{iInterest}$ represents the user's interest in the i-th topic, and T'_{ui} represents the user's interest topic distribution.

3.2.3 Citation Distribution Let $\{d_1, d_2, d_3, \ldots, d_i\}$ represent the set of academic resources read by a user within a period. Establish a citation relationship graph, and the user's citation set is represented by $C_u = (C_{u1}, C_{u2}, C_{u3}, \ldots, C_{un})$.

4 Precise Recommendation

4.1 Similarity Matching

By extracting text features of academic resources, we establish the academic resource model $M_d = \{K_d, T_d, C_d\}$ from three dimensions: feature words, topic words, and citations. Combined with user interest measurement, we establish the user interest model $M_u = \{K_u, T_u, C_u\}$ based on the academic resource model.

We use Jaccard similarity to calculate the similarity between user feature word preference K_u and academic resource feature word distribution K_d, as shown in Formula (7):

Formula (7)

We use cosine similarity to calculate the similarity between user topic preference T_u and academic resource topic distribution T_d, as shown in Formula (8):

Formula (8)

Based on the citation structure graph, we use the SimRank algorithm to calculate citation structure similarity. SimRank recursively defines similarity, with constant $c \in (0, 1)$ as the damping factor. Initial assignment is as follows:

Formula (9)

where $I(C)$ represents the set of adjacent points pointing to C. If $I(C_{ui})$ or $I(C_{dj})$ is empty, then $sim_{l+1} = (C_{ui}, C_{dj}) = 0$.

Formula (8) represents the citation structure similarity between vertices d_i and d_j in the citation structure graph. Formula (8) is called recursively until values converge, with the final convergence value being the citation structure similarity between academic resources C_{ui} and C_{dj}.

Define the user's interest value UID as the similarity between user interest model M_u and academic resource model M_d, calculated as shown in Formula (10):

Formula (10)

where $r_1 + r_2 + r_3 = 1$, with specific weights allocated according to experimental training.

The TOP-N academic resources with the highest user interest value UID are recommended to users.

4.2 Recommendation Algorithm Process

As shown in Figure 2 [Figure 2: see original paper], the precise recommendation algorithm proposed in this paper follows this process: (1) Obtain academic resources through web crawler tools; (2) Extract academic resource information (resource ID, title, abstract, keywords, citations, etc.) and establish citation structure network graphs; (3) Preprocess extracted academic resource information (word segmentation, stop word removal, etc.); (4) Calculate feature word distribution, LDA topic distribution, and citation similarity for each academic resource to build the academic resource model; (5) Record user behavior (browsing time, downloading, forwarding, collecting, etc.) based on Web logs to calculate interest in browsed academic resources; (6) Build user interest models based on user interest and academic resource models; (7) Calculate similarity between user interest models and academic resource models to obtain interest values for each academic resource; (8) Recommend the TOP-N academic resources with highest interest values to users.

5 Experiments

5.1 Data Collection and Processing

The experimental data source is the China Academic Journals (Online Edition) CAJ-N database. We selected papers from 2007-2018 in Chinese journals (55) in the field of library and information science and digital libraries as experimental data. After removing special issues, conference notices, and incomplete papers, we obtained 10,227 valid papers, crawling titles, abstracts, keywords, and citations. Before segmentation, we imported keywords of experimental papers, "Library and Information Science Dictionary," and "Chinese Thesaurus" as segmentation dictionaries into the Chinese Academy of Sciences NLPIR Chinese word segmentation system, and established synonym and stop word tables to improve segmentation effects. We performed word segmentation and stop word removal on titles, abstracts, and keywords of academic resources, then conducted word frequency statistics on feature words, calculated TF-IDF values, and selected the top 5 nouns or verbs with highest TF-IDF values as feature words. We vectorized texts as document-feature word matrices to build feature word distribution models of academic resources. In the LDA modeling process, we used the Gibbs sampling method in MCMC for parameter estimation, with topic number $K = 50$, document-topic hyperparameter $\alpha = 0.2$, topic-word distribution parameter $\beta = 0.01$, and Gibbs sampling iterations set to 1000. Through LDA-Gibbs model training, we obtained document-topic distributions for 10,227 documents and word distributions for K topics. Partial topic words and keyword distributions are shown in Table 1 .

Using UCINET software, we constructed the citation network of experimental papers as shown in Figure 3 [Figure 3: see original paper], where nodes represent scientific literature and directional lines between nodes indicate citation relationships, revealing associations between documents through citation relationships. We calculated vertex similarity using the SimRank algorithm.

5.2 Experimental Settings

To verify the accuracy of the proposed precise recommendation algorithm, we invited 30 library and information science students as experimental subjects. Each user performed at least 20 searches in the library and information science directory based on their interests or tasks to ensure sufficient user behavior data. Various behavior data including browsing time, searching, collecting, downloading, forwarding, scrolling, and page turning were obtained through embedded JavaScript code. During the experiment, the entire user behavior dataset was divided into two parts: 80% as training set to generate user interest models, and 20% retained as test set to verify algorithm recommendation effectiveness. In user interest model construction, combined with the established resource content model, we calculated each experimental subject's UID interest value using the algorithm proposed in this paper, setting $r_1 = 0.3$, $r_2 = 0.4$, $r_3 = 0.2$. We recommended resources with the highest TOP-5, TOP-10, TOP-15, and TOP-

20 user interest values to users, with 10 recommendations total. After each recommendation, users accessed resources they were interested in.

5.3 Results Evaluation

5.3.1 Recommendation Effectiveness Evaluation Metrics To evaluate the recommendation effectiveness of the constructed model, this paper selected precision, recall, and F-value as evaluation metrics, calculated as shown in Formula (10):

Formula (10)

where A represents the number of recommended interesting resources, B represents the number of recommended uninteresting resources, and C represents the number of interesting resources not recommended.

5.3.2 Comparative Experiment Debugging This experiment selected content-based (LDA topic model) recommendation algorithm and user-based collaborative filtering recommendation algorithm for comparison. In comparative experiments using the LDA topic model, the topic number K needs to be set. Table 2 shows the impact of different K values on precision, recall, and F-value when recommending 20 academic resources per user, indicating that $K = 20$ yields the best recommendation effect.

When using user-based collaborative filtering algorithm for recommendation, different numbers of nearest neighbors produce different recommendation effects. Table 3 shows the impact of different nearest neighbor numbers h on precision, recall, and F-value when recommending 20 academic resources per user, indicating that 30 nearest neighbors yield the best recommendation effect.

5.3.3 Results Comparison Based on comparative experiment debugging results, with topic number set to 20 and nearest neighbor number set to 30, content-based recommendation algorithm and user-based collaborative filtering algorithm achieve optimal recommendation effects. Under these experimental conditions, we calculated precision, recall, F-value, and average precision for three algorithms under different recommendation numbers. Experimental results are shown in Figures 4-7 [Figure 4: see original paper][Figure 5: see original paper][Figure 6: see original paper][Figure 7: see original paper].

Experimental results show that as recommendation numbers increase from 5 to 20, precision decreases sequentially while recall and F-measure increase sequentially. With the same recommendation number, the user interest measurement-based recommendation algorithm proposed in this paper achieves the highest precision and F-value, demonstrating the best recommendation effect, followed by collaborative filtering algorithm, and finally content-based recommendation

algorithm. Across the entire experiment, the average precision of the user interest measurement-based collaborative filtering algorithm is 14% higher than collaborative filtering algorithm and 23% higher than content-based recommendation algorithm. This shows that considering user behavior and citation relationships makes the algorithm proposed in this paper better at predicting user interests with superior recommendation effects.

Conclusion

Knowledge discovery systems provide a data foundation for precise recommendation services through their rich resource data and user data. Through fragmentation, fine-grained mining, and analysis of data resources, discovery systems can deeply present content features of resources, reveal semantic relationships, establish citation associations, achieve deep resource aggregation, discover hidden connections between resources for users, reveal new knowledge patterns, and provide refined knowledge discovery services. As user needs become fragmented, refined, and personalized, discovery systems must fully utilize user behavior data, measure user interests, understand user needs, provide precise knowledge recommendation services for users, enhance user interaction experience, meet users' knowledge needs, and promote multiplication of knowledge value. Precise recommendation is an important function for enhancing digital library knowledge discovery service capabilities, bringing innovative growth points to digital library knowledge discovery services.

From the perspective of user interest, this paper extracts resource content features from three dimensions—feature words, topic words, and citations—to establish academic resource models. Through user interest measurement, it constructs user interest models and optimizes precise recommendation for knowledge discovery services using similarity algorithms, with empirical verification of algorithm feasibility through comparison with traditional content-based and collaborative filtering recommendation algorithms. The recommendation algorithm proposed in this paper has three advantages: (1) It considers citation relationships, more scientifically revealing internal connections between academic resources. (2) It introduces user behavior sets to analyze user interest preference degrees, making recommendation results more accurate and objective. (3) When user interests change, the recommendation algorithm can capture recent interest changes to recommend more suitable information. The algorithm can grasp user interests in real-time for precise recommendation, improving discovery system knowledge service capabilities and enhancing user experience. This paper also has limitations, such as cumbersome algorithm steps, large computational load, limited experimental samples and time, and some manual control in the experimental process with certain subjectivity. Therefore, in future research, we will further improve algorithm performance, simplify algorithm steps, enhance algorithm applicability, and increase recommendation accuracy.

References

[1] Bi Qiang, Liu Jian. Research on digital literature resource aggregation and service recommendation methods based on domain ontology[J]. Journal of the China Society for Scientific and Technical Information, 2017, 36(5): 452-460. [2] WALTERS W H. Google Scholar coverage of a multidisciplinary field[J]. Information Processing & Management, 2007, 43(4): 1121-1132. [3] YANG S Q, WAGNER K. Evaluating and comparing discovery tools: how close are we toward next generation catalog?[J]. Library hi tech, 2010, 28(4): 690-709. [4] MICHAEL G. The evaluation of discovery services at Lynchburg College: 2009-2010[J]. College & undergraduate libraries, 2012, 19(2-4): 387-397. [5] Qin Hong. Discussion on digital resource perception service framework in pervasive computing environment[J]. Library and Information Service, 2014, 58(5): 13-16, 21. [6] Zhang Jun. Research on library knowledge discovery service based on user portrait[J]. Library and Information, 2017(6): 60-63. [7] GUAN P, WANG Y F. Personalized scientific literature recommendation based on user's research interest[C]//International conference on natural computation, Fuzzy systems and knowledge discovery. Changsha: IEEE, 2016: 1273-1277. [8] RICCI F, ROKACH L, SHAPIRA B, et al. Recommender systems handbook[M]. New York: Springer, 2011. [9] RAZMERITA L. An ontology-based framework for modeling user behavior: a case study in knowledge management[J]. IEEE transactions on systems, man, and cybernetics-part A: systems and humans, 2011, 41(4): 772-783. [10] Li Xueming, Li Hairui, Xue Liang, et al. TFIDF algorithm based on information gain and information entropy[J]. Computer Engineering, 2012, 38(8): 37-40. [11] Wang Zhenzhen, He Ming, Du Yongping. Text similarity calculation based on LDA topic model[J]. Computer Science, 2013, 40(12): 229-232. [12] Wang Chuanqing, Bi Qiang. Research on deep aggregation of digital resources from hypernetwork perspective[J]. Journal of the China Society for Scientific and Technical Information, 2015(1): 4-13. [13] Liu Hongwei, Gao Hongming, Chen Li, et al. Interest identification management model based on user browsing behavior[J]. Data Analysis and Knowledge Discovery, 2018(2): 74-85. [14] Zeng Ziming, Jin Peng. Research on digital library knowledge recommendation service based on user interest change[J]. Library Tribune, 2016, 36(1): 94-99. [15] KRISHNAMOORTHY R, SUNEETHA K R. User interest estimation using behavior monitoring measure[J]. Transplantation, 2013, 78(2): 651-652. [16] CLAYPOOL M, BROWN D, LE P, et al. Inferring user interest[J]. IEEE internet computing, 2001, 5(6): 32-39. [17] ZHENG L, CUI S, YUE D, et al. User interest modeling based on browsing behavior[C]//International conference on advanced computer theory and engineering. Chengdu: IEEE, 2010: V5-455-V5-458. [18] Zhang Haipeng. Research on personalized recommendation based on Web log mining[D]. Chongqing: Chongqing University, 2007. [19] JEH G, WIDOM J. SimRank: a measure of structural-context similarity[C]//Eighth ACM SIGKDD international conference on knowledge discovery and data mining. Edmonton: ACM, 2002: 538-543. [20] Yin Liling, Liu Baisong, Wang Yangyang. Research on cross-type high-quality academic resource recommendation algorithm[J]. Journal of the China Society for Scientific and Technical

Information, 2017, 36(7): 715-722.

Author Contributions

Ding Mengxiao: Designed research plan and wrote paper;
Bi Qiang: Proposed research ideas and revised paper;
Xu Pengcheng: Data collection and experiments;
Li Jie: Refined research ideas and revised paper;
Mu Dongmei: Refined research ideas and revised paper.

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