Postprint: A Method for Identifying Scholarly Academic Expertise Based on Publication Characteristics

Authors: Chen Chong, Li Nan, Liang Bing, Wang Chenlin, Xu Zeng and Xu Lin, Zheng Tingting

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

Abstract

[Purpose/Significance] Identifying scholars' academic expertise based on publication characteristics is a crucial task in scholar profiling, which holds significant value for applications such as scholar classification, review expert selection, and discovering small peer groups. [Method/Process] First, we analyze the factors that reveal academic expertise and construct an expertise label weight allocation model using the Analytic Hierarchy Process. We employ TextRank and concept linking techniques to identify subject terms from Chinese and English publication content, and combine weights to screen for vocabulary that reflects domain consensus and expertise summarization as expertise labels. Researchers from multiple fields who have obtained talent titles are selected, expertise labels are extracted from their representative Chinese or English publications, and the expertise areas listed in talent announcements are used as a reference benchmark to evaluate the identification effectiveness through manual scoring and semantic computation. [Results/Conclusion] Among scholars tagged with Chinese expertise labels, 71.9% of individuals' expertise descriptions were deemed satisfactory. Among scholars tagged with English expertise labels, 77.2% of individuals' expertise descriptions were deemed satisfactory. The experiments demonstrate that the proposed method for identifying scholars' academic expertise is reasonable. The main innovations are: proposing solutions for mining candidate label words from literature content under conditions of different languages (Chinese and English) and whether external knowledge bases exist; combining bibliometric factors, using multiple publication characteristics to screen expertise labels, and proposing a weight allocation scheme; and addressing the lack of evaluation benchmarks, proposing a supplementary approach based on semantic computation, thereby expanding evaluation methods.

Full Text

Preamble

Vol. 63 No. 20 October 2019

Identifying Scholars' Academic Expertise Based on Publication Features

Chen Chong[1], Li Nan[1], Liang Bing[2], Wang Chenlin[1], Xu Zengxulin[1], Zheng Tingting[1]
[1]School of Government Management, Beijing Normal University, Beijing 100875
[2]Institute of Scientific and Technical Information of China, Beijing 100038

Abstract

[**Purpose/Significance**] Identifying scholars' academic expertise based on publication features is a critical task in scholar profiling, with significant value for scholar classification, reviewer selection, and discovering peer experts. [**Method/Process**] This study first analyzed factors that reveal academic expertise and constructed an Analytic Hierarchy Process (AHP) model for allocating weights to expertise tags. Using TextRank and conceptual linking techniques, topical terms were identified from Chinese and English publication content. Combined with weight-based screening, vocabulary with domain consensus and expertise representativeness was selected as expertise tags. Researchers with prestigious titles from multiple domains were selected, and expertise tags were extracted from their representative Chinese or English publications. Using the expertise fields listed in talent announcements as the benchmark, identification effectiveness was evaluated through manual scoring and semantic computation. [**Result/Conclusion**] Among scholars tagged with Chinese expertise labels, 71.9% were considered satisfactory. For English expertise labels, 77.2% were considered satisfactory. Experiments demonstrate the proposed method's 合理性. Key innovations include: (1) proposing solutions to mine candidate tags from publication content under different conditions (Chinese vs. English, availability of external knowledge bases); (2) combining bibliometric factors to screen expertise tags using multiple publication features with a proposed weight allocation scheme; and (3) addressing the lack of evaluation benchmarks by supplementing evaluation methods with semantic computation.

Keywords: scholar profiling, expertise tagging, analytic hierarchy process, term extraction, expertise tagging evaluation

Classification Number: G250.7

1 Introduction

As human society advances toward a knowledge economy, individuals who master knowledge become increasingly valuable resources. Scholars represent a typi-

cal category of such resources, characterized by rich attributes [1], among which academic expertise is most crucial for characterizing their knowledge profile. Academic expertise can be used for scholar classification and retrieval, helps discover peer experts [2] to facilitate collaboration, and enables more accurate selection of paper or project reviewers [3-4]. Additionally, applications targeting knowledge communities—such as literature retrieval systems, personalized learning, and collaborative knowledge platforms—require improved service precision based on users' academic expertise. Therefore, identifying scholars' academic expertise is a research problem of significant practical importance.

In this paper, academic expertise refers to research directions in which scholars excel. Existing methods for acquiring expertise fall into two categories: (1) direct extraction from personal homepages or resumes, and (2) identification of appropriate vocabulary from publication content. In the first approach, scholars typically provide only a small number of expertise descriptors with inconsistent expression habits. Research shows that only 21.3% of scholars list research interests on their homepages [5], making the completeness, standardization, and timeliness of such extractions unreliable.

The second approach expands the source and quantity of expertise tags from publication content, enabling more objective and comprehensive labeling while allowing timely discovery of new interest tags. However, a key challenge is how to automatically identify vocabulary that simultaneously possesses domain consensus and expertise representativeness—particularly when dealing with scholars across multiple domains and publications in both Chinese and English.

The core problems in characterizing scholars' expertise based on publication features are twofold: (1) analyzing factors that reflect scholars' expertise, and (2) identifying vocabulary with domain consensus and expertise summarization capability as tags. For convenience, we refer to highly generalizable and well-standardized vocabulary from publication content as *topical terms*, which serve as candidate terms for *expertise tags* that identify scholars' expertise.

We first analyze important factors in scholars' publications that reveal their academic expertise, such as scholars' contributions to publications, publications' contributions to academia, and term quality for summarizing publications. Second, we employ TextRank and conceptual linking techniques to identify topical terms from Chinese and English publications, respectively. Finally, we construct a weight allocation model using AHP to score candidate terms based on multiple factors and select high-scoring vocabulary as expertise tags, thereby solving the expertise identification problem.

2 Related Research

Characterizing scholars' expertise constitutes an important aspect of scholar profiling, which analyzes scholars' personal information, publications, or academic behaviors to identify and extract appropriate tags summarizing personal characteristics, research interests, and academic influence [1]. While user profiling

in information environments aims to provide personalized services [6], scholar profiling improves precision and personalization for research communities, with expertise identification at its core.

2.1 Expertise Feature Analysis

Information reflecting scholars' expertise includes publications [7], research projects [8], collaboration networks [9], and citations [10]. Two types of information are particularly important: publication content features and scholars' contribution to publications.

2.1.1 Publication Content Features In papers and projects, author-provided keywords are closely related to research themes but focus on indexing publications rather than expressing authors' expertise. Moreover, author keywords commonly suffer from non-standardization issues, such as inconsistent semantic granularity [11], inappropriate indexing depth, and excessive use of generic terms [12-13]. Therefore, directly using author keywords to identify expertise is suboptimal and may introduce noise. Given that publication content contains richer information, mining vocabulary with strong generalizability and domain consensus offers a better approach. Related research typically identifies important terms from titles, abstracts [14], paragraph beginnings, and endings [15].

2.1.2 Scholars' Contribution to Publications Collaborative research is ubiquitous in modern science. Statistics show that in 2015, co-authored papers accounted for 92.3% of all domestic scientific papers [16]. Analysis of four core journals in library and information science reveals that 2-3 author collaborations dominate, while collaborations with four or more authors represent the future trend [17]. Since collaborators contribute differently to a publication, the same publication holds varying value in revealing each author's expertise. Therefore, selecting representative publications that reflect substantive contributions is necessary.

In research evaluation, author contribution is often empirically distinguished by author order [18]. Survey data show that approximately 82% of respondents believe author order correlates with contribution, with earlier authors contributing more [19]. According to conventions in most disciplines, first and corresponding authors are considered more closely related to publication contributions [20]. Additionally, publication volume [21], publication date, and citation count [22] reflect scholars' domain activity and publication quality, making them valuable references for determining contribution.

2.2 Topical Term Identification from Publications

A fundamental task in identifying expertise tags from publications is discovering domain-consensus topical terms. Methods can be categorized as direct or indirect based on dependence on external knowledge.

2.2.1 Direct Methods These methods measure term importance within content. Implementation approaches include constructing term co-occurrence networks and identifying high-importance nodes using algorithms like TextRank [24-26], considering term position influence [25], network coverage capability [27], or combining term frequency and position features. External features like comments, citations, and links can also calculate keyword importance, analogous to blog content tagging [28]. Semantic computation methods such as topic modeling identify topic and term distributions, using high-probability topics and terms as expertise indicators [29,30]. However, such results lack interpretability for humans, as topic semantics are implicitly expressed through word distributions, and high-probability topics only indicate frequently occurring themes, not necessarily domain expertise. With word embeddings, Word2vec has been used to construct term vectors, improving semantic computation flexibility and enhancing domain keyword identification accuracy by calculating term similarity [31].

2.2.2 Indirect Methods These methods map important vocabulary to standardized domain term spaces using external knowledge bases like domain-specific dictionaries [32] or ontologies [33]. Conceptual linking techniques [34-36] are typically employed. Using Wikipedia as an example, article titles serve as concept terms. After identifying important vocabulary through NLP methods, statistical information and associative relationships are used to automatically map terms to Wikipedia concepts, achieving standardization. Such methods apply to term normalization and semantic disambiguation, with major tools including TagMe [37] and Wikifier [38].

2.3 Differences from Related Research

Several prior studies relate to our task. Liu et al. [13] used normalized paper keywords with overlapping K-means clustering to identify expertise categories in big data but did not explore automatic generation of appropriate expertise tags. Mao et al. [39] used high-frequency nouns from full texts to construct expert graphs in computational linguistics but ignored varying contributions among co-authors. Fan et al. [40] extracted research preference tags from content and bibliometric features, adjusting weights by author contribution and publication recency, but only implemented and tested their method on two selected scholars.

Our study differs in four aspects: (1) We propose automatic identification of domain-consensus, appropriately-grained topical terms from publication content as candidate expertise tags, ensuring objectivity, richness, diversity, and timely discovery of new tags compared to direct extraction from resumes. (2) We incorporate scholars' contribution to publications as a weighting factor for candidate tags, distinguishing expertise differences among co-authors in the prevalent collaborative research environment. (3) We employ AHP to screen expertise tags, combining qualitative and quantitative approaches for computationally simple and interpretable results. (4) Facing the lack of standard benchmarks, we design

experiments combining manual scoring and semantic computation to maximize evaluation methods.

3 Expertise Tag Identification Method

Our fundamental assumption is that scholars' recent representative publications hold significant value for reflecting expertise. We can identify domain-consensus, standardized vocabulary from recent important publications where scholars made substantial contributions. Using AHP, we construct a weight allocation model for important factors reflecting expertise, determine feature importance through weight analysis, then calculate scores for all candidate topical terms from publications using direct and indirect methods, and finally select high-scoring terms as expertise tags. Since open, high-quality Chinese knowledge bases across multiple domains are scarce, while English has widely-recognized, comprehensive resources like Wikipedia, we adopt direct methods for Chinese and indirect methods for English.

3.1 Tag Weight Allocation Model

Let scholar r's publication set be D and expertise tag set be A. Tags identifying r's expertise should satisfy: (1) reflect r's research domain characteristics with generalizability and standardization (denoted as B_1); and (2) originate from important publications where r made substantial contributions, reflecting r's contribution to the expertise domain (denoted as B_2).

Based on our investigation in Section 2.1, we define the expertise tag weight allocation hierarchical model shown in [Figure 1: see original paper]. B_1 primarily includes original keywords C_1 and candidate topical terms C_2. B_2 mainly includes author order C_3 and citation count C_4, representing r's contribution to publications and the academic community's recognition.

We use AHP to calculate feature weights for expertise tags. AHP is a quantitative-qualitative method that decomposes decision problems into hierarchical structures, obtains priority weights of lower-level elements to upper-level elements by solving judgment matrix eigenvectors, then aggregates layer-by-layer to obtain final weights for the overall goal, using the maximum final weight as the optimal solution. This method offers interpretable results and has been applied in scholar influence evaluation [41] and academic value assessment [42].

We used the Delphi method to determine pairwise relative importance among factors in [Figure 1: see original paper] and construct judgment matrix M. M represents the importance comparison result between lower-level factors i and j for an upper-level factor, as shown in formula (1):

$$M_{ii} = 1, \quad M_{ij} = \quad (i, j = 1, 2) \quad \text{formula (1)}$$

Regarding factors B_1 and B_2, compared to vocabulary generalizability and standardization, r's contribution to the expertise domain more significantly impacts expertise tag identification. We consider B_2 slightly more important than B_1, thus determining the $A \rightarrow B$ judgment matrix.

For factors C_1 and C_2, original keywords C_1 are author-provided and naturally important for expressing content and themes, with many studies using keywords for topic mining and expertise identification. C_2 represents automatically identified terms from content. C_1 is considered more recognized than C_2, making C_1 slightly more important than C_2, thus determining the $B_1 \rightarrow C$ judgment matrix M_1.

For factors C_3 and C_4, given the prevalence of co-authorship, C_3 reflects author order jointly recognized by all authors, indicating r's substantive contribution. C_4 represents the publication's domain contribution. Even if r's publication has high citations, a low author order (non-corresponding) suggests limited contribution. Therefore, C_3 is significantly more important than C_4 for reflecting r's substantive contribution, thus determining the $B_2 \rightarrow C$ judgment matrix M_2.

All three judgment matrices pass consistency tests, indicating no self-contradiction in relative importance. Finally, we obtain matrix M's eigenvector $W = (0.22, 0.11, 0.54, 0.13)$. Table 1 shows the weight coefficients:

Table 1 Hierarchical Indicators and Weights

Scholar Expertise Tag A	Vocabulary Generalizability & Standardization B_1	0.33	Scholar's Domain Contribution B_2	0.67
Original Keywords C_1	0.22			
Candidate Topical Terms C_2	0.11			
Author Order C_3			0.54	
Citation Count C_4			0.13	

3.2 Topical Term Identification

For scholar r's publication document set $D = \{d_1, d_2, \ldots, d\}$, the identified domain topical term set is denoted as $T = \{t_1, t_2, \ldots, t\}$. This section implements the task $D \rightarrow T$: identifying topical terms from publication content.

3.2.1 Chinese Topical Term Identification Due to the lack of open, multi-domain Chinese knowledge bases, we adopt direct methods to generate T. First,

we construct a term co-occurrence network for D and calculate term node importance. TextRank [24] is a common method for measuring node importance in co-occurrence networks, using a voting mechanism where nodes vote for neighbors within a given co-occurrence window threshold, with vote weight depending on the node's own received votes. A node's TextRank value is calculated from neighbor votes, and sorting by TextRank yields candidate topical terms.

3.2.2 English Topical Term Identification We adopt indirect methods to identify T from English publications. Using Wikipedia article titles as concepts, conceptual linking techniques [37] calculate mapping probability and consistency to map important vocabulary to controlled concepts. This avoids synonymy and ambiguity issues in expertise identification. [Figure 2: see original paper] shows an example of TagMe annotating topical terms in an English abstract, mapping "machine learning techniques" to the Wikipedia concept "Machine Learning" as a candidate topical term.

3.3 Expertise Tag Selection

In [Figure 1: see original paper], the stage from C_1, C_2 to B_1 selects expertise tag set T' from scholar r's topical term set $T = \{t_1, t_2, \ldots, t\}$ and author keyword set T, implementing task $(T, T) \rightarrow T'$. For each term t T, we calculate corresponding weights. If t T, then $t^1 = 1$ and $t^2 = 0$; otherwise, $t^1 = 0$ and $t^2 = 1$.

The stage from C_3, C_4 to B_2: Let m denote author order position, with $t^3 = 1/m$, treating corresponding authors as first author. Let n denote citation count of t's publication, and n' denote total citations of r's all publications, then $t^4 = n/n'$.

Finally, formula (2) calculates each topical term t's score as a candidate expertise tag. We compare scores of all candidate tags for scholar r, sum scores for identical tags, and select tags with scores above a threshold or top-ranked tags as r's expertise tags. In our experiments, we selected the top 5 tags.

$$s(t) = 0.22 \times t_{C_1} + 0.11 \times t_{C_2} + 0.54 \times t_{C_3} + 0.13 \times t_{C_4} \quad \text{formula (2)}$$

4 Experiments and Results

4.1 Data and Method

4.1.1 Experimental Data We selected 557 researchers from the "Beijing Science and Technology New Star" talent program list (2013-2017) published on the Beijing Municipal Science and Technology Commission website, covering science, engineering, agriculture, forestry, medicine, and other domains. We collected their publications and projects from the past five years from Baidu Scholar and personal homepages, including titles, keywords, and abstracts. For each author, we sorted papers by citation count in descending order and selected

three representative works: if Chinese papers were majority in top five, we selected three Chinese papers; otherwise, three English papers. If fewer than five publications were available, we selected two works from the majority language. After removing researchers with insufficient data, we obtained 480 researchers (set R), including 300 with Chinese publications (R , 867 papers) and 180 with English publications (R , 520 papers).

4.1.2 Identification Method For Chinese publications, we preprocessed titles and abstracts using the HanLP toolkit, then applied TextRank combined with frequency and length heuristics to extract important terms as topical terms. For English publications, we used TagMe to map abstract vocabulary to Wikipedia concepts. After obtaining candidate topical terms, we calculated scores using AHP and selected the top 5 terms as expertise tags.

4.2 Experimental Results

Tables 2 and 3 show sample expertise tag identification results for Chinese and English publications. S represents the professional field from "Science and Technology New Star" applications, serving as our evaluation benchmark. T is the author keyword set from high-impact representative works. T is the candidate topical term set extracted from representative works. T' is the final expertise tag set obtained by our method.

Visually, extracted topical terms T partially overlap with T but have less fine-grained semantics, expressing broader themes. T exhibits uneven granularity—for example, in r_1, "epidural anesthesia" in T is more specific than "patient anesthesia" in T , while in r_4, "Quality of service" in T is overly generic.

Table 2 Sample Chinese Expertise Tag Identification Results

Researcher	Benchmark S	Author Keywords T	Topical Terms T	Expertise Tags T'
r_1	Clinical Anesthesia	Propofol; Propofol dosage; Isopropofol; Isoflurane; Dosage; General anesthesia; Epidural anesthesia…	Propofol; Propofol dosage; Isopropofol; Isoflurane; Patient anesthesia; Colon cancer…	Propofol; Isopropofol; Isoflurane; Propofol dosage; Patient anesthesia

Researcher	Benchmark S	Author Keywords T	Topical Terms T	Expertise Tags T'
r_2	Urban Rail Transit	Urban rail transit; Rail transit industry structure; Life cycle; Integrated management…	Urban rail transit; Urban track; Life cycle; Intelligent management; Integrated management…	Urban rail transit; Rail transit; Rail transit industry chain; Life cycle; Integrated management

Table 3 Sample English Expertise Tag Identification Results

Researcher	Benchmark S	Author Keywords T	Topical Terms T	Expertise Tags T'
r_3	Neurosurgery	Ganglioglioma; Hydrocephalus; Ventricular system; Surgical pathology; Von Hippel-Lindau disease; Hemangioblastoma…	Ganglioglioma; Hydrocephalus; Ventricular system; Von Hippel-Lindau disease; Hemangioblastoma…	Ganglioglioma; Hydrocephalus; Ventricular system; Von Hippel-Lindau disease; Surgical pathology
r_4	Computer Software	Quality of service; Service-oriented architecture; Computational modeling; Basel problem; Cloud computing; Web services; Petri nets…	Service-oriented architecture; Big Data; Web services; Petri nets; Computational modeling; Interval arithmetic…	Service-oriented architecture; Big Data; Web services; Petri nets; Computational modeling

5 Evaluation and Analysis

A major challenge in similar studies is the absence of recognized standard evaluation sets. We used the professional fields from Science and Technology New Star announcements as evaluation set S. However, S contains few terms with coarse granularity, making it an imperfect benchmark. Therefore, our evaluation consists of two parts: (1) manual scoring, and (2) semantic similarity calculation between T' and S using word vectors. To ensure credibility, we first tested evaluators' domain identification abilities and verified scoring consistency.

5.1 Evaluator Domain Identification Ability Test

This test analyzed volunteers' understanding of professional domains and ability to identify domain expertise. We randomly selected 50 researchers from R and R each, presenting their abstracts and keywords in multiple-choice format with four similar domain terms (including the correct answer). Volunteers selected the most relevant domain based on content understanding. Options were generated using Word2vec similarity to the standard answer. For example, in Table 4 , the standard answer "Chemical Power Supply" in S was supplemented with three semantically similar options. Correct selections were scored accordingly.

We recruited 10 volunteers from engineering, science, and social science disciplines. Those achieving 90% accuracy (9 for Chinese, 4 for English) qualified as evaluators for manual scoring.

Table 4 Domain Identification Ability Test Example

Read the following abstract and keywords, then select the author's most relevant research domain:

Abstract: Based on battery appearance, electrical performance, environmental adaptability, and safety testing items, considering potential hazards including fire, explosion, gas/liquid leakage, noise, vibration, machinery, and electricity during testing, this paper proposes corresponding safety protection requirements and suggestions from personnel, sample, equipment, and environmental perspectives.

Keywords: battery testing laboratory safety protection

Options: A. Chemistry B. Materials Chemistry **C. Chemical Power Supply** D. Environmental Chemistry

5.2 Manual Evaluation of Expertise Tags

Evaluators assessed the reasonableness of generated expertise tags based on domain knowledge and evaluation standards. For each r R , we provided Chinese expertise tags and representative Chinese works; for r R , English tags and works. Scores were: Satisfactory (1), Uncertain (0), or Unsatisfactory (-1). "Satisfactory" indicated tags that reflect expertise, have reasonable semantic granularity, and possess domain consensus.

Using SPSS, we calculated Kendall's W coefficient for inter-evaluator consistency. Chinese evaluation achieved 0.924 consistency; English achieved 0.921. At $\alpha = 0.05$, all $p < 0.001$, indicating highly significant consistency.

Following majority rule, if "1" scores reached half or more of evaluators, the automatically generated tags for r were deemed satisfactory. Satisfactory ratios P_1 were 65.3% for R and 68.3% for R .

5.3 Semantic Computation Supplementary Evaluation

Due to S's limited quantity and coarse granularity, we conducted supplementary semantic evaluation on researcher set R' (those not scoring "1"). We trained Word2vec on all publication titles, keywords, abstracts, and professional fields, preprocessing texts through segmentation, POS tagging, stopword removal, and retaining nouns, verbs, and adjectives. Each term's vector was averaged from constituent element vectors.

We calculated cosine similarity between T' and S. If a researcher had half or more tags with similarity ≥ 0.9, the result was deemed satisfactory. Satisfactory ratios P_2 were 6.6% for R' and 8.9% for R'.

5.4 Experimental Result Analysis

Combining both evaluation phases, overall satisfactory ratios P reached 71.9% for Chinese and 77.2% for English expertise tags. Given S's limitations in quantity and granularity, these accuracy rates demonstrate our method's reasonableness.

Our approach overcomes the limitations of inconsistent granularity and excessive fine-grained terms in literature keywords, using appropriately generalizable vocabulary to represent expertise. Evaluation benchmarks from talent announcements are often overly coarse—for example, r_1's benchmark "Clinical Anesthesia" prevents literal matching confirmation. We adopted semantic proximity principles for evaluation, as shown in Table 5 .

Table 5 Analysis of Satisfactory Researcher Ratios

Language	Manual Scoring P_1	Semantic Computation P_2	Overall P
Chinese	65.3%	6.6%	71.9%
English	68.3%	8.9%	77.2%

Manual evaluation faced cognitive barriers for highly specialized domains, while semantic computation reduced subjective judgment errors. We also noted that corpus scale limited word vector computation, preventing some domain terms and tags from being calculated. Overall, our method effectively identifies Chinese and English expertise tags with generalizability across domains.

6 Conclusion and Future Work

We analyzed important factors revealing academic expertise in scholars' publications, constructed an AHP model for weight allocation, and used TextRank and conceptual linking to identify topical terms from Chinese and English publications, providing rich candidate terms for expertise description. By integrating multiple weighting factors, we selected domain-consensus, expertise-summarizing tags, solving the expertise identification problem. Evaluation

through manual scoring and semantic computation demonstrates our method's reasonableness, with satisfactory rates of 71.9% (Chinese) and 77.2% (English) on test data.

Key contributions include: (1) Proposing and implementing large-scale expertise tag identification that integrates publication content and academic contributions through AHP, generating uniformly-grained, domain-consensus tags. (2) Exploring direct and indirect topical term generation methods for different languages and knowledge base availability, broadening applicability. (3) Proposing a reasonable experimental design combining manual and computational evaluation to supplement limited benchmarks.

Future improvements include: (1) Expanding experimental data scale for more comprehensive evaluation. (2) Refining experimental design by incorporating peer review opinions and author contribution statements to optimize weight allocation, as some domains don't follow contribution-based authorship. (3) Developing universal methods for mixed Chinese-English texts, as Chinese publications may contain English terminology that cannot be strictly separated by language.

References

[1] Yuan S, Tang J, Gu X. A survey on scholar profiling techniques in the open internet[J]. Journal of Computer Research and Development, 2018, 55(9): 1903-1919.

[2] Zhu W, Li C. Empirical research on peer expert selection based on concept knowledge networks[J]. Journal of Intelligence, 2017, 36(7): 78-83, 88.

[3] Zhao L, Feng S, Liu T, et al. How to select "small peer" reviewers[J]. Acta Editologica, 2007, 19(1): 75.

[4] Cheng X. Research on peer review expert identification for scientific projects[D]. Beijing: Institute of Scientific and Technical Information of China, 2016.

[5] Tang J, Yao L, Zhang D, et al. A combination approach to web user profiling[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), New York: ACM, 2010, 5(1): 1-44.

[6] Hu Y, Mao N. Construction of digital library knowledge community user model based on user profiling[J]. Library Theory and Practice, 2017(4): 82-85.

[7] Gong J, Liu L. Expert recommendation based on personal knowledge maps[J]. Chinese Journal of Management, 2011, 8(9): 1365-1371.

[8] Tang J, Zhang D, Yao L. Social network extraction of academic researchers[C]//Seventh IEEE International Conference on Data Mining. Piscataway: IEEE, 2007: 292-301.

[9] Yan M, Yu Z, Zhang Y, et al. An expert recommendation approach combining project correlation and professional ability[C]//International Conference on Fuzzy Systems and Knowledge Discovery. Piscataway: IEEE, 2015: 1220-1224.

[10] Zhang S, Liang M, Cao G. Research on citation-based scientific literature theme extraction[J]. Information Studies: Theory & Application, 2017, 40(6): 122-127.

[11] Deng Q, Wang J. Common problems and analysis in keyword indexing[J]. Science-Technology & Publication, 1999(2): 36.

[12] Wang S. Analysis of keyword indexing quality in Chinese academic journals[J]. Journal of Yan'an University (Social Sciences Edition), 2001, 23(3): 97-99.

[13] Liu X, Zhu D, Wang X, et al. Research on multi-expertise expert identification: A case study in big data[J]. Library and Information Service, 2018, 62(3): 55-63.

[14] Ren H, Wang D, Wang F. Research on the relationship between keyword combination novelty and paper academic impact[J]. Library and Information Service, 2017, 61(9): 87-93.

[15] Zhao Y, Guo G. Information retrieval algorithm based on iterative theme word extraction[J]. Journal of South China University of Technology (Natural Science Edition), 2004(S1): 77-80.

[16] Yu Z, Jia J. Analysis of co-authorship in Chinese scientific papers[J]. Global Science, Technology and Economy Outlook, 2017, 32(Z1): 92-100.

[17] Zou D. Analysis of co-authorship in four amphibious core journals in library and information science[J]. Journal of Agricultural Library and Information Science, 2016, 28(3): 61-64.

[18] Cui L, Lu Y. Analysis of author contribution elements based on author order: Taking author contribution statements in *Library and Information Service* 2015-2016 as examples[J]. Library and Information Service, 2017, 61(9): 80-86.

[19] Zuo J. Research on author order and contribution in scientific co-authorship[D]. Chongqing: Southwest University, 2014.

[20] Jia X, Wang X, Li Z, et al. Authorship issues for equal-contribution authors and co-corresponding authors in scientific papers[J]. Chinese Journal of Scientific and Technical Periodicals, 2012, 23(4): 603-605.

[21] Li Z. Comprehensive evaluation of core authors using publication and citation counts: Taking *Journal of Agricultural Library and Information Science* as an example[J]. Journal of Agricultural Library and Information Science, 2007, 19(10): 161-163.

[22] Li P, Zhou J, Yang G. Citation analysis of papers in *Journal of the China Society for Scientific and Technical Information* based on CSSCI[J]. Library and

Information Studies, 2009, 2(2): 48-52.

[23] Xie R, Li X, Han X, et al. Construction of author influence evaluation index based on weighted citation frequency and author order[J]. Information Science, 2018, 36(8): 90-93, 111.

[24] Mihalcea R, Tarau P. TextRank: Bringing order into texts[C]//Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2004: 404-411.

[25] Xia T. Research on keyword extraction using position-weighted TextRank[J]. New Technology of Library and Information Service, 2013, 29(9): 30-34.

[26] Liu Z, Huang W, Zheng Y, et al. Automatic keyphrase extraction via topic decomposition[C]//Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2010: 366-376.

[27] Liu P, Zhou M. Expertise mining based on co-word networks[J]. Information Science, 2012, 30(12): 1815-1819.

[28] Tsai T, Shih C, Peng T, et al. Explore the possibility of utilizing blog semantic analysis for domain expert and opinion mining[C]//Conference on Intelligent Networking and Collaborative Systems. Piscataway: IEEE, 2009: 241-244.

[29] Zhang X, Lu W, Cheng Q. Application of PLSA in expertise identification for library and information science[J]. New Technology of Library and Information Service, 2012, 28(2): 76-81.

[30] Du Y, Zhang W, Liu T. User interest identification based on topic-enhanced convolutional neural networks[J]. Journal of Computer Research and Development, 2018, 55(1): 188-197.

[31] Ning J, Liu J. Research on keyword extraction integrating Word2vec and TextRank[J]. New Technology of Library and Information Service, 2016, 32(6): 20-27.

[32] Lu W, Liu J, Qin X. Expert retrieval and evaluation in library and information science based on expertise vocabulary[J]. Journal of Library Science in China, 2010, 36(2): 70-76.

[33] Hu Y, Liu P. Research on expertise representation based on ontology concepts[J]. Library and Information Service, 2012, 56(4): 17-21, 40.

[34] Lu W, Wu C. A survey on entity linking[J]. Journal of the China Society for Scientific and Technical Information, 2015, 34(1): 105-112.

[35] Mihalcea R. Wikify!: Linking documents to encyclopedic knowledge[C]//Conference on Information and Knowledge Management. New York: ACM, 2007: 233-242.

[36] Luo P. Research on document organization methods based on concept hierarchy[D]. Beijing: Beijing Normal University, 2014.

[37] Ferragina P, Scaiella U. TagMe: On-the-fly annotation of short text fragments (by Wikipedia entities)[C]//Conference on Information and Knowledge Management. New York: ACM, 2010: 1625-1628.

[38] Tsai C, Roth D. Illinois cross-lingual Wikifier: Grounding entities in many languages to the English Wikipedia[C]//International Conference on Computational Linguistics. New York: ICCL, 2016: 146-150.

[39] Mao J, Li G. A method for constructing researcher profiles in research fields based on OKM[J]. Library and Information Service, 2014, 58(14): 34-40.

[40] Fan X, Dou Y, Zhao P, et al. Research on constructing researcher profiles integrating multi-source data[J]. Library and Information Service, 2018, 62(15): 31-40.

[41] Du J, Zhang B, Tang X. Dual measurement of author academic influence: Integrating citation influence and collaboration influence[J]. Journal of the China Society for Scientific and Technical Information, 2014, 33(4): 388-395.

[42] Pan Q, Wu C, Cheng J. Research on academic value evaluation of scientific papers based on fuzzy AHP theory[J]. Acta Editologica, 2001, 13(1): 16-18.

Author Contributions

Chen Chong: Research design, paper writing
Li Nan: Research design, evaluation, paper writing
Liang Bing: Research design and evaluation methodology refinement
Wang Chenlin: English tag identification experiments, weight algorithm design
Xu Zengxulin: Chinese tag identification experiments
Zheng Tingting: Literature review

Abstract: [Purpose/significance] Identifying expertise tags of scholars is the most critical task in scholar profiling. Expertise tags contribute to finding peer experts, clustering domain scholars and selecting reviewers. [Method/process] This study analyzed related factors on the scholar expertise in academic publications, then constructed a hierarchical analysis model on the weight allocation of the factors. The TextRank algorithm has been used to identify topical terms in Chinese corpus, and the conceptual linking technique in English corpus. The extracted terms, together with the previously analyzed factors have been combined to select the expertise tags of the scholars. In this study, a group of honored scholars of different domains have been selected. Their research expertise information from their resumes have been taken as evaluation benchmark. And the expertise tags extracted from their publications have been compared with the benchmark by human judgment and additional semantic similarity judgment. [Result/conclusion] The evaluation shows that the expertise tags of 71.9% scholars are acceptable for Chinese, and 77.2% for English. The experiment proves that the method proposed in this article is pragmatic and may lead to reasonable

results. The chief innovation of this study lies in three aspects. Firstly, term extraction approaches that suit to different application conditions have been explored, such as the language of publication and the availability of domain knowledge base. Secondly, multiple features have been combined together to identify the expertise tags of scholars, including the content of publications, the substantial contribution to the publications of the scholars, and the influence to the domain of the publications. Thirdly, a reasonable experimental design and evaluation method is proposed, and the proposed approach has been verified by combining manual scoring and semantic calculation results.

Keywords: scholar profiling, expertise tagging, analytic hierarchy process, term extraction, evaluation on expertise tagging

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