International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(REVIEW ARTICLE)

Check for updates

Advancements and characteristics of scholar profiling research in China

Qiu Xiaohua *

School of International Trade and Economics, Central University of Finance and Economics, China.

International Journal of Science and Research Archive, 2024, 12(01), 787–794

Publication history: Received on 12 April 2024; revised on 18 May 2024; accepted on 21 May 2024

Article DOI: https://doi.org/10.30574/ijsra.2024.12.1.0881

Abstract

This study analyzes the development and characteristics of scholar profiling research in China from 2016 to 2022, utilizing content analysis of 38 core journal articles. Scholar profiling, adapted from user profiling in big data, involves extracting and analyzing academic data to build models that describe scholars' characteristics and behaviors. Our findings highlight the use of diverse data sources, including commercial databases and social platforms, and emphasize the importance of data preprocessing and modeling techniques. While applied research dominates, the field faces challenges such as the lack of standardized data collection and depth in model construction. The study underscores the need for robust profiling tools and broader application scenarios to enhance the effectiveness of scholar profiling systems.

Keywords: Scholar Profiling; China; Characteristics; Advancements

1. Introduction

In the era of big data, the volume of data has been increasing significantly, characterized by its variety, complexity, and the accelerating rate of updates. Industries are saturated with vast amounts of data. The critical challenge in the big data era is extracting valuable information from this massive, heterogeneous, multi-source, and high-frequency data[1]. User profiling has been recognized as a tool for effectively utilizing big data and has been widely applied. Internet companies like JD.com, Alibaba, ByteDance, and Baidu profile users based on their interaction data, categorize them, and provide tailored services to meet the specific characteristics of different user groups, achieving a personalized experience for each individual[2][3].

The concept of user profiling was first introduced by Alan Cooper[4], the father of interaction design, in 2006 under the term "User Persona," which represents a virtual model of a real user based on actual data. Related concepts include[5][6] "User Portrait," "Customer Segment," and "User Persona," as described in Table 1. The term "User Profile" refers to a model that outlines user needs based on data, originating from the big data domain.

Table 1 Concepts related to user profiling

Term	Meaning	Source Domain
User Portrait	Character Portraiture	Art
Customer Segment	User Needs Analysis	Business Models
User Persona	Collection of Users' Natural and Social Attributes	Design
User Profile	A Model Depicting User Needs Based on Data	Big Data

* Corresponding author: Qiu Xiaohua

At its core, user profiling is a framework that describes a user's interests, characteristics, behaviors, and preferences[7]. Essentially, user profiling is a labeled representation of a user's profile, constructed through the classification of user attributes and extraction of features using specific technical methods, culminating in a complete user profile[8].

With the development of big data technologies, the scope and application scenarios of user profiling have expanded, and the concept has gained widespread acceptance and use. Scholar profiling is a structured description of researchers' behaviors, characteristics, and research needs based on bibliometric data, representing an application of user profiling technology in the field of scientific research[9].

This paper aims to provide an overview of domestic research on scholar profiles, exploring the progress in this area, including the construction models, data sources, and application fields of scholar profiles. We conducted a content analysis of domestically published papers of China related to scholar profiling to offer references for future research in this field.

2. Fundamentals of Scholar Profiling

2.1. Definition of Scholar Profiling

Scholar profiling, mirroring the definition and scope of user profiling, focuses on researchers and concentrates on research data. It involves extracting relevant data sets from extensive research data, classifying, clustering, and visualizing these sets to discover basic information, research behaviors, and interests of scholars, and ultimately constructing a scholar model through multidimensional tags.

2.2. Construction Process of Scholar Profiling

The essence of scholar profiling is the process from researcher data to researcher tags, and finally to user profiles. The construction process can be divided into five steps: data collection, data preprocessing, statistical modeling, mining modeling, and application scenarios. Each step includes specific considerations and focuses.

Data collection is a crucial phase in constructing scholar profiles, where the breadth and validity of data sources directly affect subsequent analyses. It involves aspects of data sources and collection methods. Data preprocessing includes data normalization, name disambiguation, and named entity recognition. Statistical modeling involves analyzing demographic characteristics and research outputs of scholars, while mining modeling uses data mining techniques to reveal internal features of scholars, such as interests, collaborations, academic influence, mentor-mentee relationships, and team dynamics. The final phase, application scenarios, includes recommendations for collaborators, knowledge recommendations, and visualizations.

3. Research Design

This study employs a content analysis method, a systematic coding process to describe and interpret textual resources. The key aspects of content analysis include selecting samples, coding, and elaborating on the coding results.

3.1. Data Sources

For this study, we utilized the China National Knowledge Infrastructure (CNKI) and Wanfang databases to search for documents using the keywords "Scientist Profiling" or "Scholar Profiling" or "Academic Profiling." The search was conducted on November 1, 2023, yielding 112 results. After preliminary screening to remove irrelevant documents, 77 papers were retained, including 19 master's theses and 58 journal articles. Given the comprehensive nature of the master's theses, this paper focuses primarily on the journal articles. After thoroughly reading these 58 papers, 38 were found to be closely related to scholar profiling research. These 38 papers were subjected to content analysis.

The trends indicate that domestic research on scholar profiling began to increase around 2016, peaking in 2020. Based on current trends, it is expected that the number of papers on scholar profiling will continue to grow.

3.2. Coding Framework

The coding framework is fundamental for content analysis. This study uses a three-category framework for coding: type of literature, content theme, and application research. The categories reflect the scope and distribution of scholar profiling research in China, as well as the methodologies and characteristics of scholar profiling systems.

The categories of literature type are shown in Table 2:

- Model Building Research: Analysis and research on the dimensions of scholar profiles without empirical studies.
- Applied Research: Research based on existing scholar profile models, which collect data, extract relevant information, construct scholar profiles, and apply these models in practical scenarios.
- Reviews and Evaluations: Comprehensive reviews and critiques of existing scholar profiling research.

Table 2 Categories of Literature Types

Primary Category	**Explanation**	**Coding**
Model Construction	Proposing a new model for scholar profiling or revising an existing model	A1
Applied Research	First propose a model for scholar profiling, then collect data, extract relevant data, construct the scholar profile, and finally apply the model in specific scenarios.	A2
Review and Critique	Literature Review and Critique	A3

Content themes are primarily based on the keywords of the papers. Keywords are extracted, statistically analyzed, and summarized to form the content themes shown in Table 3.

Table 3 Content Theme Categories

Research Topic	**Coding**	**Quantity**
University Library	B1	5
Academic Journal	B2	2
Multi-source Heterogeneity	B3	4
Academic Influence	B4	3
Recommendation System	B5	3
Tagging System	B6	10
Research Interests	B7	8
Data Mining	B8	7
Visualization	B9	10
Academic Social Network	B10	3
Academic New Media	B11	4
Precision Services	B12	4

The application research categories are developed based on the scholar profiling process, encompassing five primary categories and several subcategories, as shown in Table 4.

3.3. Reliability of Coding

The validity of the category system and the consistency of coding results are crucial for ensuring the reliability of content analysis. Although due to time constraints, this aspect was not fully developed in this paper.

Table 4 Application Research Categories

Primary Category	Secondary Category	Tertiary Category	Coding	Quantity
Data Collection	Data Source	Commercial Database	C1	10
		Social Platform	C2	4
		Internet Data	C3	6
		Log Data	C4	1
	Collection Method	Data Scraper	D1	9
		Retrieval Export	D2	6
		Program Interface	D3	1
Data Preprocessing	Data Normalization		E1	4
	Name Disambiguation		E2	4
	Named Entity Recognition		E3	12
Statistical Modeling	Demographic Characteristics	Name, Employer, Job Title, Honors, Awards	F1	6
	Research Outputs		F2	4
Mining Modeling	Scholar Interests		F3	6
	Scholar Collaboration	Degree Centrality	F4	3
		Closeness Centrality	F5	2
		Betweenness Centrality	F6	2
	Academic Influence	PageRank Value	F7	2
		H-Index	F8	4
		Citation Frequency	F9	2
	Mentorship Relationships		F10	2
	Team Mining		F11	7
Application Scenarios	Collaborator Recommendation		G1	3
	Knowledge Recommendation		G2	3
	Visualization	Word Cloud	G3	7
		Statistical Dashboard	G4	2
		Knowledge Graph	G5	4
		Other Graphics	G6	3

4. Results and Analysis

4.1. Analysis of Literature Type Categories

The distribution of the types of literature is shown in Table 5. Reviews and evaluation papers account for 24%, applied research papers for 50%, and model building research papers for 26%. This distribution reflects the practical application of scholar profiles, which constitutes a significant portion of the literature, highlighting the practical implications and advantages of scholar profiling in this applied field.

Table 5 Statistical Distribution of Literature Types

Literature Types	Quantity	Sorting
applied research	19	1
model building	10	2
Reviews and evaluation	9	3

4.2. Analysis of Research Themes

As illustrated in Table 6, domestic research on scholar profiling primarily focuses on label systems and visualization, accounting for 15.9% each. This is followed by research interests (12.7%) and data mining (11.1%). Other significant themes include academic libraries (7.9%), multi-source heterogeneity (6.3%), academic new media (6.3%), and precision services (6.3%). These themes indicate common application areas, with academic libraries being a frequent application scenario, reflecting the practical implementation of big data technologies to profile researchers and provide targeted services. Label systems and visualization are highly focused areas within scholar profiling systems, indicating current hot topics in this field.

Table 6 Distribution of Research Themes

Research Themes	Quantity	Sorting
Tagging System	10	1
Visualization	10	1
Research Interests	8	3
Data Mining	7	4
University Library	5	5
Multi-source Heterogeneity	4	6
Academic New Media	4	6
Precision Services	4	6
Academic Influence	3	9
Recommendation System	3	9
Academic Social Network	3	9
Academic Journal	2	12

4.3. Analysis of Application Research Categories

4.3.1. Data Collection

Data collection is the first step in scholar profiling, involving aspects of data sources and collection methods. Existing literature primarily categorizes data sources into four types: commercial databases (such as CNKI, Wanfang, Elsevier Science Direct, Springer Link, Web of Science, etc.), social platforms (like ResearchGate, Twitter, Weibo, Zhihu, etc.), internet data (such as Wikipedia, Baidu Baike, and researcher homepages), and log data (such as database search logs). Commercial databases are used in 10 papers, indicating that combining existing bibliographic data with internet data is a main approach in current scholar profiling research in China.

Different data collection methods are employed depending on the source characteristics, with common methods including web scraping, database export, and programmatic APIs. The use of web scraping is highlighted in 9 papers, accounting for 47% of the literature analyzed, indicating that scholar profiling in China heavily relies on this method due to the lack of structured and standardized research data.

4.3.2. Data Preprocessing

Data collected from various sources cannot be used directly for profiling and requires preprocessing, which includes data normalization, name disambiguation, and named entity recognition. Twelve papers address named entity recognition, while four papers each discuss data normalization and name disambiguation.

Data Normalization: Data integration involves consolidating data from multiple sources into a single database[10]. Due to different metadata standards and the lack of a unified description of researcher information across databases, it is necessary to normalize fields that carry the same meaning but are represented differently across sources.

Name Disambiguation: The issue of name disambiguation includes homonymy (different people with the same name) and synonymy (the same person known by different names)[11]. This has been extensively researched and various methods have been developed, including rule-based, machine learning-based, semantic fingerprint-based, and unique identifier-based approaches.[12]

Named Entity Recognition: This process identifies all person names, organization names, geographical locations, as well as dates, currencies, and percentages mentioned in texts.[13] The identification of organization names, personal names, and place names has been well studied, with several open-source Chinese natural language processing toolkits available for use, such as Fudan University's fastNLP and Beijing Institute of Technology's NLPIR system. After recognizing entities, it is crucial to align and merge identical entities.

4.3.3. Statistical Modeling

Statistical modeling involves calculating demographic and research output features of scholars. Demographic features include name, gender, title, employer, honors, awards, educational background, and work history. Research output features encompass journal articles, conference papers, academic monographs, patents, conference presentations, and projects. Six papers discuss the statistical analysis of demographic features, and four papers cover research outputs.

4.3.4. Mining Modeling

Mining modeling builds on the data to use data mining and artificial intelligence methods to unearth characteristics underlying scholarly behavior. The literature primarily discusses mining scholar interests, collaborative networks, academic influence, mentorship relationships, and research team dynamics. Team mining is mentioned in 7 papers, making it the most discussed application within scholar profiling, which reflects 37% of the analyzed papers.

4.3.5. Application Scenarios

Scholar profiling is applicable in various scenarios, including recommendations for research collaborators, evaluations of research outputs, and research management. In research collaboration networks, scholar profiling technology can effectively identify potential collaborators. Multi-dimensional scholar profiles allow for a more comprehensive evaluation of academic impact, moving beyond traditional metrics centered on impact factors. This analysis also includes personalized knowledge recommendation services based on scholars' interests.

In the analyzed papers, three discuss collaborator recommendations, three cover personalized knowledge services, seven utilize word clouds for visualization, two present scholar data on dashboards for a comprehensive view, four use knowledge graphs for visualization, and three employ other graphical methods to depict scholar characteristics.

5. Discussion

5.1. Characteristics of Domestic Scholar Profiling Research

The analysis of domestic scholar profiling literature reveals several characteristics of the research landscape in China:

Theoretical Descriptions and Reviews: There is a prevalence of theoretical and review literature, whereas complete systems for scholar profiling are less common. This may be due to the diversity of data sources, significant differences in data formats, and the technical challenges associated with ongoing data updates in China.

Model Building: Model construction is a focus in domestic scholar profiling research. Based on the characteristics of available data sources and application scenarios, multidimensional models have been developed, forming the theoretical foundation for scholar profiling systems.

Information Extraction and Recommendation Algorithms: These are critical aspects of application systems. During the profiling process, it is essential to extract keywords that match scholarly characteristics first. Recommendation algorithms are a major application direction for scholar profiling, hence their prominence in the current research.

Dependence on Technological Advancements: Scholar profiling relies on advancements in name disambiguation, information extraction, machine learning, visualization, and artificial intelligence, which are integral to the progress of this field. Most researchers choose mature computer science technologies, though some tailor these technologies to better fit the specific needs of scholar profiling based on its unique characteristics.

5.2. Challenges in Domestic Scholar Profiling Research

Several challenges persist in domestic scholar profiling research:

- Data Collection Standards: Many scholars use web scraping to collect data, indicating a dispersion of researcher data across various platforms without a clear structural standard. The lack of open metadata contributes to the difficulties in data collection for scholar profiling in China.
- Lack of Depth in Theoretical Model Construction: Although statistical modeling is common due to its reliance on accurate and timely data, mining models often lack depth, failing to reveal the nuanced characteristics of scholar profiling, such as the trajectory of a scholar's academic career, shifts in research interests, and modeling of scholarly search behavior.
- Absence of Mature Scholar Profiling Tools: Most scholar profiling systems are custom-built, lacking the mature tools found in bibliometrics. This suggests potential for further development and expansion in this field, especially in creating user-friendly tools based on more standardized databases.
- Limited Application Scenarios: Current scholar profiling applications are somewhat limited, often integrated within other systems rather than standing alone. Integration with document retrieval systems is particularly common, likely because scholar profiling relies heavily on bibliographic data, making it convenient to develop these systems on top of existing databases.

6. Conclusion

The study provides a comprehensive analysis of the progress and challenges in scholar profiling research in China from 2016 to 2022, highlighting the use of diverse data sources, advanced modeling techniques, and the application of artificial intelligence. Despite the achievements, significant obstacles such as the lack of standardized data collection, the need for deeper theoretical model construction, and the absence of mature tools remain. This research underscores the necessity for robust profiling tools and broader application scenarios to enhance the effectiveness of scholar profiling systems. Ultimately, this study contributes to the improvement of scholar profiling methods, benefiting the academic community and supporting future advancements in this field.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Wang D, Li Q, Zhang Z, Wang Z. Research on the construction method of scholar profiles. J Inform Sci. 2022;41(08):812-821.

[2] Dong W, Xiong H, Du J, Wang N. Research on the recommendation of scientific collaborators based on scholar profiles. Data Anal Knowl Discov. 2022 [cited 2022 Dec 01];1-19.

[3] Shi X, Liu P. Review on the identification of scholars' research interests. Data Anal Knowl Discov. 2022;6(04):16-27.

[4] Li Y, Liu Z, Gao Y. Research on a multilingual author topic model for scholar interest profiling. J Inform Sci. 2020;39(06):601-608.

[5] Mo J, Dou Y, Kai Q. Construction of research team profiles based on multi-source heterogeneous data. Theory Pract Inform. 2020;43(09):100-106.

[6] Zhang L, Zhang X, Wu Y, Guo S. Research on the construction of dynamic user profile models for social academic apps based on small data. Libr Inform Work. 2020;64(05):50-59.

[7] Yuan R, Wang Q. Construction and empirical study of academic blog user profile models: A case study of ScienceNet blogs. Libr Inform Work. 2019;63(22):13-20.

[8] Zhang Y, Huang J, Wang G. Method for constructing a precise stereoscopic profile of researchers' behaviors considering global and local information. J Inform Sci. 2019;38(10):1012-1021.

[9] Chen C, Li N, Liang B, Wang C, Xu Z, Zheng T. Method for identifying scholars' academic expertise based on the characteristics of their outputs. Libr Inform Work. 2019;63(20):96-103.

[10] Jia T, Xia F. Research on the behavior of researchers from the perspective of scientometrics. Big Data. 2019;5(05):38-47.

[11] Peng C, Wu B. Research on the scholar profiling system in "smart campus". Digit Libr Forum. 2019;(02):2-11.

[12] Yuan S, Tang J, Gu X. Review of scholar profiling technology in the open internet. Comput Res Dev. 2018;55(09):1903-1919.

[13] Fan X, Dou Y, Zhao P, Zhou X. Research on the construction method of researcher profiles by integrating multi-source data. Libr Inform Work. 2018;62(15):31-40.