Journal of Infrastructure, Policy and Development 2024, 8(11), 7281.
https://doi.org/10.24294/jipd.v8i11.7281

Article

An unsupervised machine learning-based profile system of Chinese researchers

Yan Yu†, Peiyu Xu†, Shuo Liu†, Taiming He†, Lu Yang, Junqiang Zhang*

School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu 611731, Sichuan, China
*** Corresponding author:** Junqiang Zhang, zhangjq@uestc.edu.cn
† The first four authors contributed equally to this article.

CITATION

Yu Y, Xu P, Liu S, et al. (2024). An unsupervised machine learning-based profile system of Chinese researchers. Journal of Infrastructure, Policy and Development. 8(11): 7281.
https://doi.org/10.24294/jipd.v8i11.7281

ARTICLE INFO

Received: 21 June 2024
Accepted: 27 August 2024
Available online: 15 October 2024

Abstract: The construction of researcher profiles is crucial for modern research management and talent assessment. Given the decentralized nature of researcher information and evaluation challenges, we propose a profile system for Chinese researchers based on unsupervised machine learning and algorithms. This system builds comprehensive profiles based on researchers' basic and behavior information dimensions. It employs Selenium and Web Crawler for real-time data retrieval from academic platforms, utilizes TF-IDF and BERT for expertise recognition, DTM for academic dynamics, and K-means clustering for profiling. The experimental results demonstrate that these methods are capable of more accurately mining the academic expertise of researchers and performing domain clustering scoring, thereby providing a scientific basis for the selection and academic evaluation of research talents. This interactive analysis system aims to provide an intuitive platform for profile construction and analysis.

Keywords: researcher profiles; machine learning; unsupervised learning; expertise recognition; cluster scoring

1. Introduction

In the era of big data, user profiling has become an important tool widely used in various industries. Researchers are vital to scientific and technological progress, as their skills directly impact research quality and efficiency. Therefore, understanding and evaluating their abilities is crucial for effective management, talent development, and collaboration. Researcher profiles help educational organizations and talent management departments to better understand the personal characteristics, academic background, research expertise, and academic dynamics of researchers (Chavez et al., 2023; Papaevangelou et al., 2023). The number of researchers in China today is vast, making the construction of researcher profiles highly valuable (Zhang et al., 2019). Constructing researcher profiles has both theoretical and practical significance (Jia et al., 2017). Traditional methods for evaluating researchers primarily rely on expert reviews and personal resumes, which often involve a degree of subjectivity and certain limitations. Additionally, the reliance on manual data collection and assessment can result in evaluations that may not always be objective. Furthermore, the dispersed nature of researcher information presents significant challenges in collecting, analyzing, and assessing this data, thereby greatly reducing efficiency.

With the rapid development of artificial intelligence technology, machine learning technology can help researchers process data faster and better. Bahar et al. (2017) proposed an innovative approach called 'ScholarLens' that employs a range of Natural Language Processing (NLP) techniques. Olavo Holanda et al. (2013) proposed an agent-based approach for automatically generating researcher profiles from multiple data sources and providing customized services. Hassan Noureddine et al. (2015) proposed an innovative approach called the "Context-Aware Researcher Profiling" (CARP). This method integrates multiple heterogeneous data sources and utilizes semantic network techniques to create unified and validated researcher profiles through data matching, clustering, and merging. de Campos et al. (2020) proposed a text clustering-based approach to identify experts' points of interest, which better reflects experts' interests and areas of specialization. Boussaadi et al. (2020) proposed a researcher profiling method based on Latent Dirichlet Allocation (LDA) topic modeling. This approach combines two LDA implementations, Gensim and Mallet, to construct profiles of researchers' areas of expertise by analyzing the articles they are interested in Tang (2016) established the AMiner Academic Platform, which uses a comprehensive modeling strategy to create models of papers, authors, and institutions. The platform offers knowledge retrieval and researcher relationship exploration services, helping to extract researcher characteristics and build profiles. Machine learning can effectively integrate and analyze researchers' multidimensional data, allowing for comprehensive assessment of their abilities and potential. This improves talent management.

However, due to the time-varying nature of researchers' behaviors and achievements, the majority of previous studies were analyzed in databases, which lacked a certain degree of timeliness (Wang, 2019). Besides, for labeled feature extraction of researchers, the majority of them used supervised machine learning and deep learning. Supervised learning requires a large amount of data to be obtained in advance and also requires labeling and training. The sheer volume and complexity of data pertaining to researchers necessitates a significant investment in training and debugging, a process that is inherently costly. Additionally, modern evaluation systems focus on exploring the correlation between achievements and research directions through in-depth semantic mining and computational analyses of researchers' existing information, so as to become more insightful information labels (Al-Shamri, 2016; Bulut et al., 2017). Unsupervised machine learning is well-suited for most demanding scenarios, as it excels at uncovering hidden patterns. This approach can effectively integrate and analyze researchers' multidimensional data, allowing for comprehensive assessment of their abilities and potential. This enhances talent selection, development, and evaluation, improves project-team matching, boosts collaboration efficiency, and refines resource allocation. Cluster analysis of expertise and achievements also reveals new research areas and competitive dynamics, benefiting academic institutions, HR departments, companies, and government agencies needing to identify research experts quickly.

For the above consideration, we used Python to design a profile system for Chinese researchers based on the unsupervised machine learning.

The main contributions of this paper are as follows:

1) The acquisition of data in real time and through various channels can track the dynamic changes of researchers to a certain extent and is more timely;
2) Unsupervised machine learning method combined with Bidirectional Encoder Representations from Transformers (BERT) pre-training model is used to mine text information and identify expertise tag words to describe the expertise of

researchers, which is more fine-grained than the description of mainstream academic platforms;

3) To address the issue of directional changes in research topics of researchers, the Dynamic Topic Model (DTM) time-topic model is used to calculate the evolutionary relationships between research topics at different stages;
4) The K-means clustering algorithm is used to cluster and score the expertise tags of multiple researchers according to their concentration in a certain domain, which provides a reference for talent training and academic evaluation and has application value;
5) An interactive analysis system for digital profiles of researchers is designed combined with the methods in this paper. The system is capable of research personnel profile construction, domain scoring and visualization display.

2. Framework of researcher profiles system

2.1. Profile dimension definition and selection

The researcher profile is a kind of labeling model that reflects the professional characteristics and work patterns of researchers by comprehensively analyzing the basic information, research habits, research behaviors, and other aspects of the data of researchers, and then refining them (Li, 2023). It stems from the in-depth analysis of huge data sets, aiming to extract key identifiers reflecting the multidimensional attributes of individuals, and then build a detailed user description (Liu, 2018).

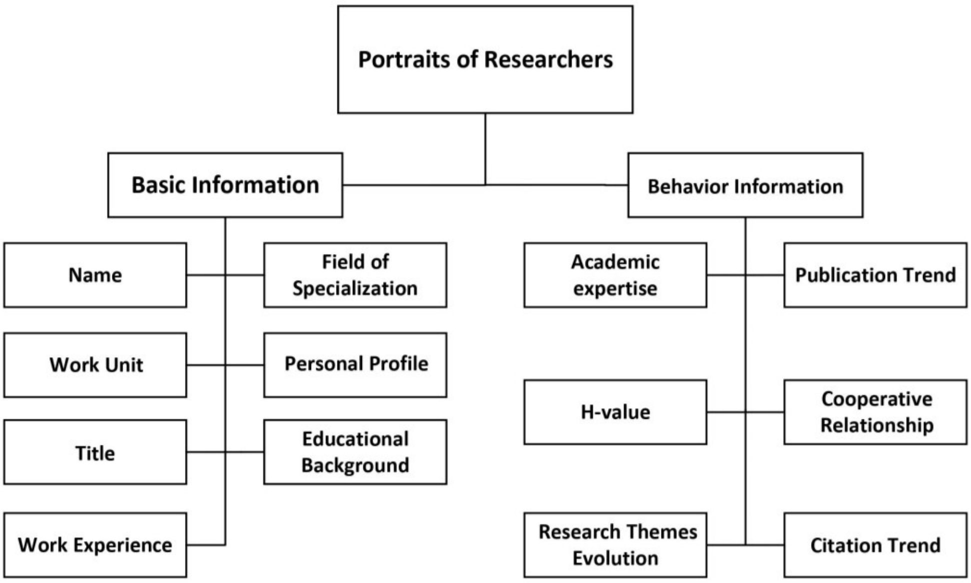


Figure 1. Schematic diagram of researcher profiles modeling.

The basic information of researchers is generally fixed and less affected by time. In contrast, the behavior information will be influenced by the changing dynamics of researchers' individual behavior over time. So, in this paper, the basic information is defined as: name, work unit, title, field of specialization, personal profile, educational background, and work experience. The research behavior information is defined as: academic expertise, H-index, research themes evolution, publication trend, collaborative relationships, and citation trend. For the application purpose of providing reference for expert discovery and academic evaluation, the basic information and research behavior information of researchers can reflect the main characteristics required. As shown in **Figure 1**, These two dimensions are modeled for researcher profiles.

Academic expertise is the most crucial of these factors, as it directly reflects the degree of focus and influence of the researcher in a particular direction (Zhao et al., 2020). For a research paper, the key information is typically summarized in the “title,” “keywords,” and “abstract” (Chamorro-Padial and Rodríguez-Sánchez, 2023; Pottier et al., 2024). The “title” directly reflects the research content of the paper; “keywords” generally provide a concise overview of the core topics and focus; and the “abstract” usually includes the research objectives, methods, results, and conclusions. Thus, extracting expertise from a researcher’s “title,” “keywords,” and “abstract” not only reveals the necessary information but also saves time and effort in processing large volumes of text data. The annual trend of citations can be used to assess the impact of a researcher’s specific research results in a given time period, while publication frequency can serve as an indicator of academic research activity. The evolution of research topics over time can be used to assess the degree of specialization of a researcher in a particular topic at a specific point in time, as well as the degree of similarity between different periods of research (Chen et al., 2019).

2.2. System technology route

The construction process of researcher profiles is divided into various modules, and **Figures 2** and **3** show the construction process of the entire system. According to the system construction flow chart (**Figure 2**), the technical route (**Figure 3**) of this paper is formulated.

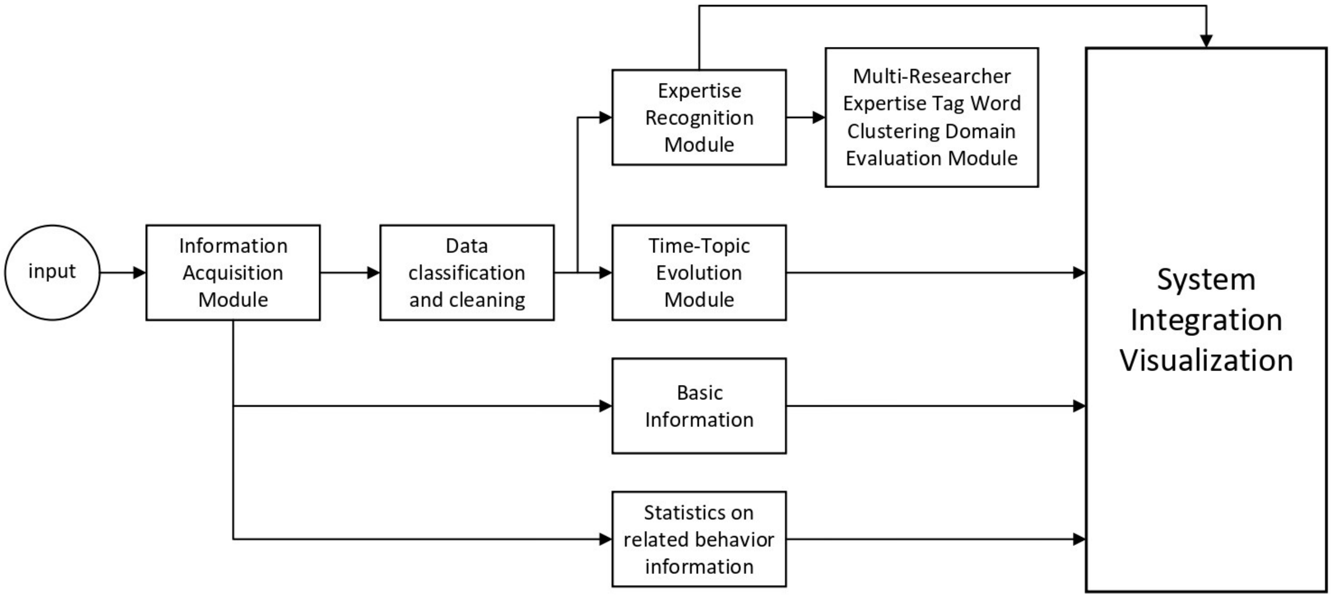

input
Information Acquisition Module
Data classification and cleaning
Expertise Recognition Module
Multi-Researcher Expertise Tag Word Clustering Domain Evaluation Module
Time-Topic Evolution Module
Basic Information
Statistics on related behavior information
System Integration Visualization

Figure 2. Researcher profiles system construction process.

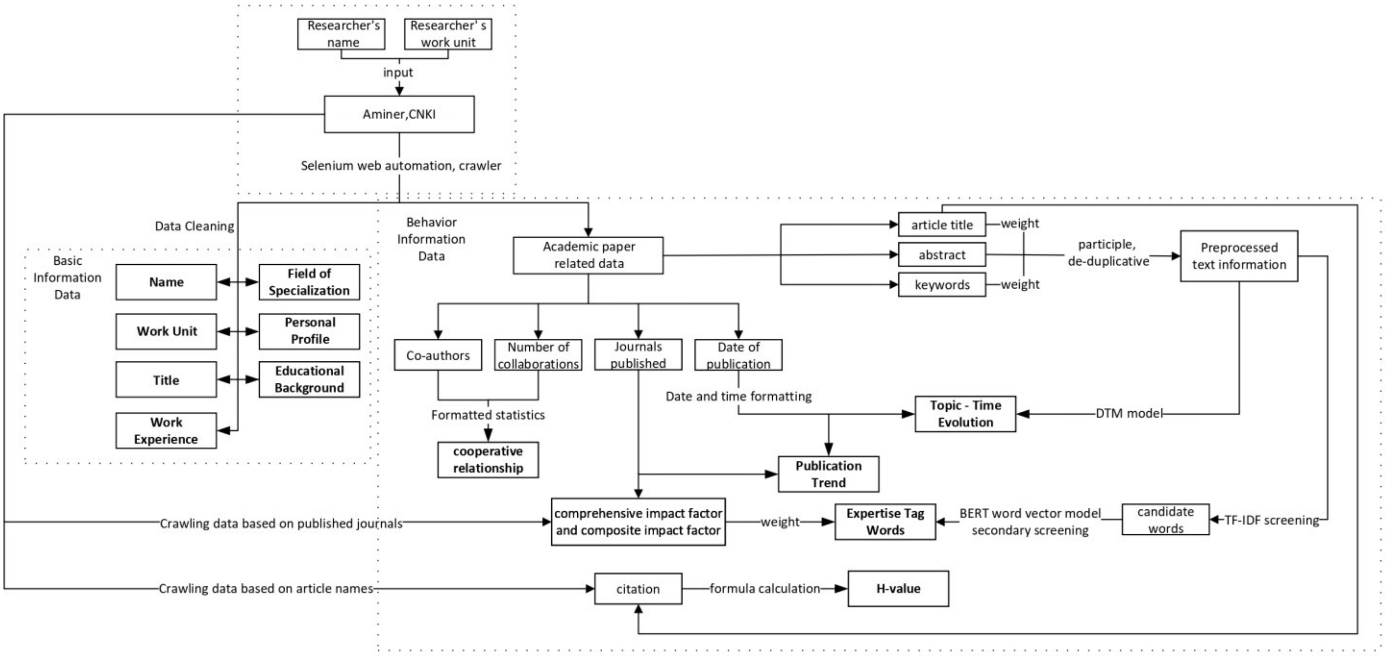


Figure 3. Researcher profiles system technical route.

AMiner is an academic science and technology information platform independently developed in China with full intellectual property rights. It offers extensive information about researchers, including their photos, titles, personal profiles, educational backgrounds, and work experiences (Tang et al., 2018). The China National Knowledge Infrastructure (CNKI) is one of the largest academic resource databases in China, covering nearly all Chinese journal papers, dissertations, conference papers, and more. As an authoritative platform for Chinese academic research, CNKI excels in both the coverage and quality of Chinese academic resources. The CNKI platform includes a wealth of multidisciplinary academic resources and related information, such as batch exports of paper details, author summaries, co-authors, citations, and various impact factors. Both AMiner and CNKI, as search-focused platforms, offer standardized information storage formats, facilitating easier access to relevant researcher data. Therefore, this paper uses AMiner to obtain basic researcher information and CNKI to acquire behavior data, selecting appropriate information from CNKI to supplement researchers' basic profiles.

First, Selenium Web automation framework and Web Crawler technology are used to collect and acquire the basic information and academic paper information related to behavior information of researchers in AMiner and National Knowledge Infrastructure (CNKI), and store them persistently. Then preprocessing operations such as word splitting and deactivation of words are carried out on the relevant information of the paper. The preprocessing results are initially screened using the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm, and then the BERT model is utilized to obtain the word vectors (Özçift et al., 2021). The secondary screening is carried out by calculating the similarity. Finally, the academic expertise of the researcher is identified with the weight value of that expertise to obtain the expertise description of the researcher. In addition, in terms of topic-time evolution, the DTM dynamic topic model is used to calculate the similarity between topics and themes between different stages to obtain the topic-time relationship evolution results. In terms of publication and citation trends, the time formatting is used to count the number of publications and citations of the researcher each year. The H-index is calculated based on the collected citation data of each paper (Hirsch, 2005; O'Leary, 2021). For the collaborative relationship, all the collected co-authors are utilized for the slice and dice process and the statistics are performed.

After building each researcher's profile, considering that the latest impact factor can more accurately reflect the current academic impact of journals and helps improve the timeliness of research assessment, academic expertise tags are weighted by latest impact factors. BERT-calculated term vectors are then clustered using K-means, and a score for each researcher is derived based on these clusters.

2.3. Interactive system

The interactive system's architecture is shown in **Figure 4**. Users log in with their account and password to access the main interface, where they enter the researcher's name and work unit. After selecting the single profile construction mode, the system collects and processes the researcher's information, saving it locally. Users can then visualize the profile. For multiple researchers, users can perform domain clustering, scoring, and visualization from the main interface (Yimam-Seid et al., 2003).

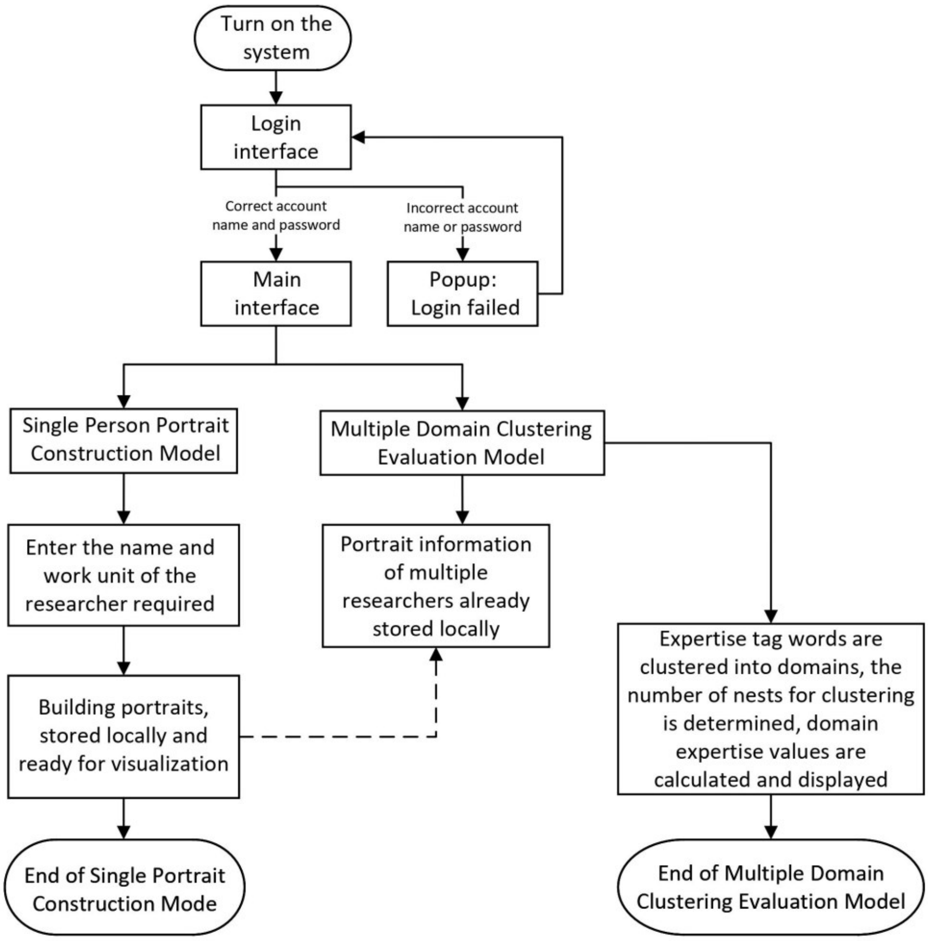

Turn on the system
Login interface
Correct account name and password
Incorrect account name or password
Main interface
Popup: Login failed
Single Person Portrait Construction Model
Multiple Domain Clustering Evaluation Model
Enter the name and work unit of the researcher required
Portrait information of multiple researchers already stored locally
Building portraits, stored locally and ready for visualization
Expertise tag words are clustered into domains, the number of nests for clustering is determined, domain expertise values are calculated and displayed
End of Single Portrait Construction Mode
End of Multiple Domain Clustering Evaluation Model

Figure 4. Interactive system workflow.

3. Main construction methods

3.1. Data acquisition and pre-processing

To obtain basic researcher information from AMiner, Selenium is used to save and reuse login cookies. By appending the researcher's name and affiliation to the search URL, data such as "title," "personal profile," "education," "work experience," and photo URLs are extracted. Photos are retrieved via GET requests, and all information is stored persistently. For behavior information from CNKI, Selenium also facilitates this process. Using developer tools, the advanced search page URL is accessed, and a script locates and exports paper details based on the author's name and affiliation. Author profiles are also collected to supplement basic information, including "title," "author," "affiliation," "source," "keywords," "abstract," and "publication date." Web Crawler technology is used to construct URLs, simulate GET requests, and merge the data for storage.

Data cleaning refers to the pre-processing of textual data, which can improve the efficiency of data processing and lay a solid foundation for subsequent information extraction and analysis. We process the 'title,' 'keywords,' and 'abstract' of each paper from researchers, weighting the 'title' and 'keywords' before merging them with the 'abstract' to form the text input data. Using suitable tokenization and stop words dictionaries, we divide the text into meaningful units and remove stop words, resulting in the final preprocessed text information for each paper.

3.2. Academic expertise recognition based on TF-IDF with BERT

The TF-IDF algorithm is used in this system to determine the candidate keywords in the text. TF-IDF assigns a weight value to each word by evaluating the frequency of a particular word within a single document and analyzing the word's popularity in the entire collection of documents, so that words that appear frequently in the current document but rarely in other documents receive a higher weight. Therefore, TF-IDF is calculated as:

$$TF\text{-}IDF(t,d,D) = TF(t,d) \times IDF(t,D) \quad (1)$$

where $TF(t,d)$ is the word frequency, which indicates how often a word appears in a document, the formula is: $TF(t,d) = \frac{n_{t,d}}{\sum_k n_{k,d}}$. $n_{t,d}$ is the number of times a word t occurs in a particular document d, while the denominator is the cumulative number of occurrences of all the words within the document d. $IDF(t,d,D)$ refers to the inverse document frequency, which indicates the generalized importance of a word in a collection of documents, and is given by: $IDF(t,D) = \log_{10} \frac{|D|}{|\{d \in D: t \in d\}|}$. $|D|$ is the total number of documents in the document collection, and $|\{d \in D: t \in d\}|$ is the number of documents containing the word t.

In determining the academic expertise tag words, the semantic contribution of the candidate words should also be evaluated. The BERT model, based on deep learning, uses the Transformer architecture and captures richer semantic information by simultaneously considering the context of a word through the self-attention mechanism (Zou et al., 2024; Xu et al., 2024). In this process, firstly, Position Embedding, Segment Embedding and Token Embedding are used to form the complete input of the BERT model; the input text is processed by a 12-layer Transformer Encoder which contains Self-Attention Mechanism and Feed-Forward Network Transformer encoder to process all the elements, and finally get the output of the word embedding vector. Each candidate word outputs 768 dimensional vectors.

Cosine similarity is used as the main similarity assessment tool to solve the problem of similarity metrics between text or data points. The formula for similarity between high dimensional word vectors is:

$$\text{Cosine Similarity}(\boldsymbol{A},\boldsymbol{B}) = \frac{\boldsymbol{A}\cdot\boldsymbol{B}}{\parallel \boldsymbol{A} \parallel \parallel \boldsymbol{B} \parallel} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{2}$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are two high-dimensional vectors; A_i and B_i are the components of vector $\boldsymbol{A}$ and vector $\boldsymbol{B}$ in the ith dimension, respectively, and n is the dimension of the vector; $\parallel \boldsymbol{A} \parallel = \sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}} A_i^2}$ is the Euclidean parameter of vector $\boldsymbol{A}$, and $\parallel \boldsymbol{B} \parallel = \sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}} B_i^2}$ is the Euclidean parameter of vector $\boldsymbol{B}$.

Inspired by Song et al. (2022) in describing expertise, the weight w_i corresponding to the academic expertise tag words are obtained using:

$$w_i = \frac{\sum_j \left(W_i + W_j\right) \times \text{Cosine Similarity}\left(V_i + V_j\right)}{w_j} \tag{3}$$

where W_i, W_j respectively denote the weights of word i and word j, and V_i and V_j respectively denote the vectors of word i and word j.

The system selects the comprehensive impact factor and the composite impact factor as the weighted factors of the academic expertise tag words. The final weight w_i' of the academic expertise tag word is calculated by $w_i' = w_i + (\text{Comprehensive Impact Factor} + \text{Composite Impact Factor})$. The academic expertise of researchers is described as:

$$\text{Academic Expertise} = \sum_i w_i' \times \text{Tag Word } i \tag{4}$$

3.3. Time-topic evolution for researchers based on DTM model

DTM is a dynamic topic model for analyzing topic changes over time intext collections and is an improvement on the Latent Dirichlet Allocation (LDA) Model by Blei et al. (2007). Traditional topic models like LDA typically assume that topics are static, whereas Dynamic Topic Model (DTM) allows topics to evolve continuously over time. DTM captures and models topic trends over time, treating topics as evolving processes by considering the chronological information of documents. The topics at each point in time can be evolved from the topics at the previous point in time through a certain probability distribution, thus capturing the evolution of topics overtime (Lei, 2017).

The preprocessed text data of researchers' academic papers are first imported into the corpus dictionary of the library. After indexing each word in the corpus, each text is converted into a two-dimensional vector using the Bag of Words (BoW) model. The corpus is analyzed in stages, and the DTM model is trained to obtain topics and the associated words for different time periods based on probability.

The Hellinger distance is a measure of the difference between two probability distributions and is often used to calculate the similarity between documents or topics. $H(P,Q)=\sqrt{\frac{1}{2}\sum_i | P(i)-Q(i)|}$, where P and Q are two probability distributions, and $P(i)$ and $Q(i)$ are the probabilities of these two distributions on the i-th event, respectively. The similarity measure was converted by calculating the Hellinger distance for each topic between stages according to $H\text{-}sim=1-H(P,Q)$. The similarity is used to reflect the evolution of each topic between stages.

3.4. Clustering domain scoring method for expertise tag words based on K-means algorithm

The system employs the K-means clustering algorithm, an unsupervised learning method (Zhao, 2009). After vectorization, the text data are represented as high-dimensional data. K-means can effectively operate in this high-dimensional space. In this paper, we use independent word vectors for each expertise tag and apply K-means to cluster these expertise tag words. This clustering helps in forming research areas and calculating scores based on the weights and similarity of the expertise tag words.

K-means algorithm needs to evaluate and select the number of clusters for clustering through metrics, in this paper, we choose to use the Inertia index and Silhouette Coefficient to evaluate the clusters (Zhu et al., 2010), and select the number of clusters for clustering through manual selection. The Inertia index is calculated as:

$$\text{Inertia}=\sum_{k=1}^{K}\sum_{x\in C_k}\| x-\mu_k \|^2 \tag{5}$$

where K is the number of clusters; C_k is the set of all data points in the k-th cluster; x is a data point; μ_k is the center of the k-th cluster; and $\| x-\mu_k \|$ denotes the Euclidean distance. The smaller the value of Inertia, the closer the data points in the cluster are, and the better the clustering effect is generally thought to be. Meanwhile, the Silhouette Coefficient is calculated as:

$$S=\frac{b-a}{\max(a,b)} \tag{6}$$

where S denotes the Silhouette Coefficient; a is the cohesion and b is the separation, for a given data point, a denotes the average distance from the point to other points within the same cluster, and b denotes the average distance from the point to all points in the nearest cluster. The higher the value of the Silhouette Coefficient, the higher the similarity between the data point and its cluster, and the lower the similarity between the data point and other clusters.

Based on the clustering results of K-means, the clustering domain k is described as having a total of j expertise tag words under the domain. Denote the clustering domain by:

$$\text{Clustering Domain } k=\sum_j w_j' \times \text{Tag Word } j \tag{7}$$

and w_j' is the weight of the corresponding expertise tag word j within the clustering domain.

Based on Equations (4) and (7), when a researcher's expertise tag word i is the same as the expertise tag word j in the clustering domain k, the number of identical expertise tag words is recorded as n. The Euclidean distance between the vector of the expertise tag word i and the vector of the cluster centers of the clustering domain k is computed using $d(P,Q)=\sqrt{\sum_{i=0}^{n}\ (p_i-q_k)^2}$. The similarity is computed in accordance with $Sim=\frac{1}{1+d(P,Q)}$, and the obtained result Sim is taken as the degree of relevance of the expertise tag word j to the clustering domain k.

Define the value of the researcher's domain expertise in the expertise tag word clustering domain k as $C\times\sum_{i=0}^{n}\ w_i'\times Sim$, where C is the amplification factor. By calculating the expertise value of each researcher in the clustered domain of the expertise tag word, the research focus status of the researcher under the domain is scored. A higher score represents a higher degree of researchers' specialization in that domain.

4. Experiments and results

4.1. Single person profile construction

The data sources for the experiment are CNKI and AMiner. All Chinese academic journals published on these platforms by required researchers are selected as behavior information data. we wrote a crawler program to get the data according to the data acquisition method in 3.1. The interactive system was designed based on the framework outlined in Section 2.3., utilizing the Tkinter library for the local interface and the Pyplot library for visualization.

The interaction system's main interface is shown in **Figure 5**. After the user inputs the researcher's name and work unit and clicks the button, the system will automatically retrieve the researcher's information.

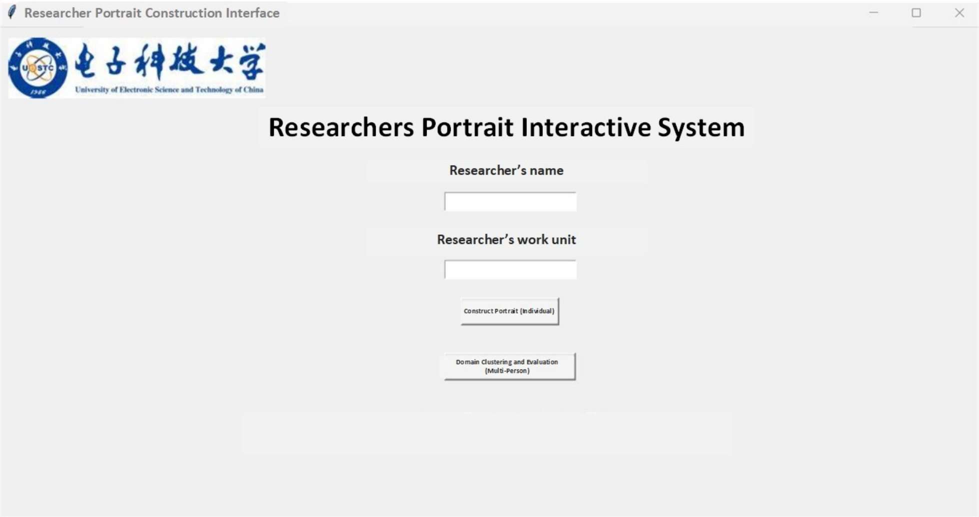

Researcher Portrait Construction Interface
电子科技大学
University of Electronic Science and Technology of China
Researchers Portrait Interactive System
Researcher's name
Researcher's work unit
Construct Portrait (Individual)
Domain Clustering and Evaluation (Multi-Person)

Figure 5. Interactive system user main interface.

We first selected 10 Chinese researchers who have published a certain number of papers in the field of library intelligence as research objects. After checking, we successfully obtained relevant information on approximately 650,000 characters, including basic information totaling about 9000 characters, and behavior information: academic papers for each researcher, along with corresponding citations, downloads, co-authors, publication dates, and the comprehensive and composite impact factors of the journals in which they were published, totaling 1425 papers. **Table 1** shows the statistics of researchers and the number of academic papers they have published.

Table 1. Acquisition of researchers' papers in the experiment.

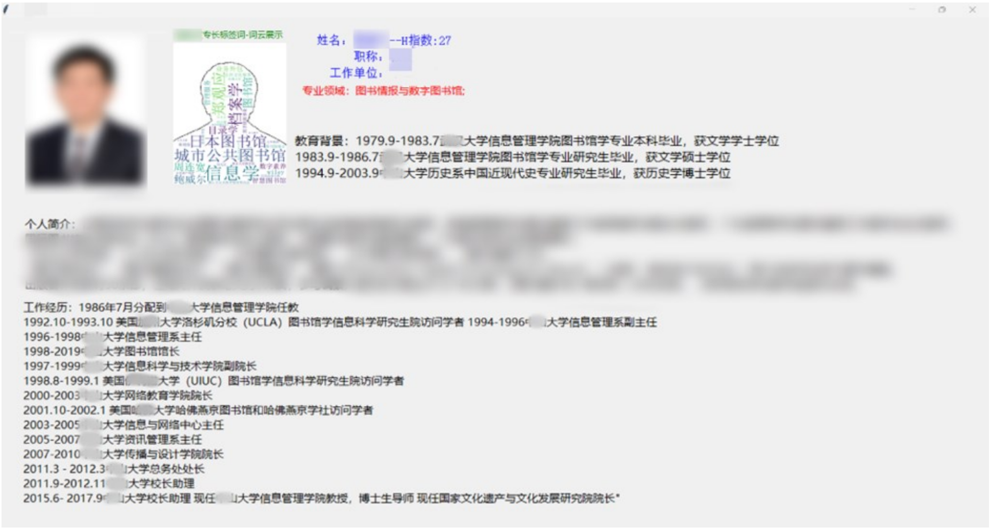
Researcher Number	Work Unit	Acquisition papers
Researcher 0	Central China Normal University	15
Researcher 1	People's Public Security University of China	56
Researcher 2	Sichuan University	59
Researcher 3	Beijing Institute of University	60
Researcher 4	Sichuan University	91
Researcher 5	Peking University	110
Researcher 6	Chongqing University	114
Researcher 7	Sun Yat-sen University	159
Researcher 8	Nankai University	358
Researcher 9	Nanjing University	403

In terms of expertise recognition, the system first preprocesses 1425 academic papers from 10 researchers for expertise recognition using the methodology described in 3.2. This includes the fusion of titles, keywords, and abstracts, where the number of titles and keywords is multiplied by 3 to weight them due to their importance. The Jieba participle tool is used and a library intelligence discipline-specific participle and deactivation thesaurus (specifically, the thesaurus list for "Discontinued Words from the Machine Intelligence Laboratory at Sichuan University" and "Harbin Institute of Technology Discontinued Words" (Gao, 2021).) is introduced to filter the text for participles and deactivated words. Then, the 10 words with the highest TF-IDF values for each paper are selected as candidate words using the scikit-learn library. The system then loads the Tensorflow_Chinese_L-12_H-768_A-12 pre-training model under the Tensor Flow framework using the encapsulated bert as serving module to train the candidate words. This pre-trained model was released by Google and is specifically optimized for Chinese natural language processing. The model's parameters were obtained through pre-training on a very large-scale Chinese corpus. This extensive pre-training enables the model to effectively capture the rich knowledge within the language, resulting in outstanding performance across various natural language processing tasks. Subsequently, the similarity calculation is performed using Equations (2) and (3). Through multiple experiments, the similarity threshold of 0.005 is determined, and secondary screening is performed. Next, weighted values are calculated using Equation (4) to obtain the description of the researcher's expertise. Finally, 50 words with the largest weighted values are selected as expertise tag words.

In terms of time-topic evolution for researchers, the system follows the method in 3.3. and uses the corpora class of the gensim library to represent the preprocessed text of the paper as a word-frequency vector, such that each word has an index and a frequency. Then, dictionaries are created to match words with word frequency vectors. We set a 5-year interval for DTM (less than 5 years were merged into the previous year) and, using Researcher 7 as an example, divided the period from 2003 to 2024 into four stages accordingly. And each phase contained 5 topics, each of which consisted of the 10 most highly weighted words. The preprocessed paper text is partitioned into 4 stages, and the Hellinger distance of each topic between adjacent time stages is calculated using the matutils class of the gensim module. On this basis, it is converted to similarity, and this is used to draw the time-topic evolution Sankey diagram.

Take Researcher 7 from Sun Yat-sen University as an example for profile construction. Once the profile is constructed, the system will pop up 3 new windows and 2 HTML pages displaying information about the researcher. To protect the privacy of researchers, all personally identifiable information in the experimental results of this paper has been anonymized or de-identified. All data is used solely for research purposes.

Figure 6 shows window 1, which contains the researcher's photo, expertise tag word cloud, name, H-index, title, work unit, area of specialization, personal profile, educational background, and work experience.



姓名：[redacted]--H指数:27
职称：[redacted]
工作单位：
专业领域：图书情报与数字图书馆;

教育背景：1979.9-1983.7[redacted]大学信息管理学院图书馆学专业本科毕业，获文学学士学位
1983.9-1986.7[redacted]大学信息管理学院图书馆学专业研究生毕业，获文学硕士学位
1994.9-2003.9[redacted]大学历史系中国近现代史专业研究生毕业，获历史学博士学位

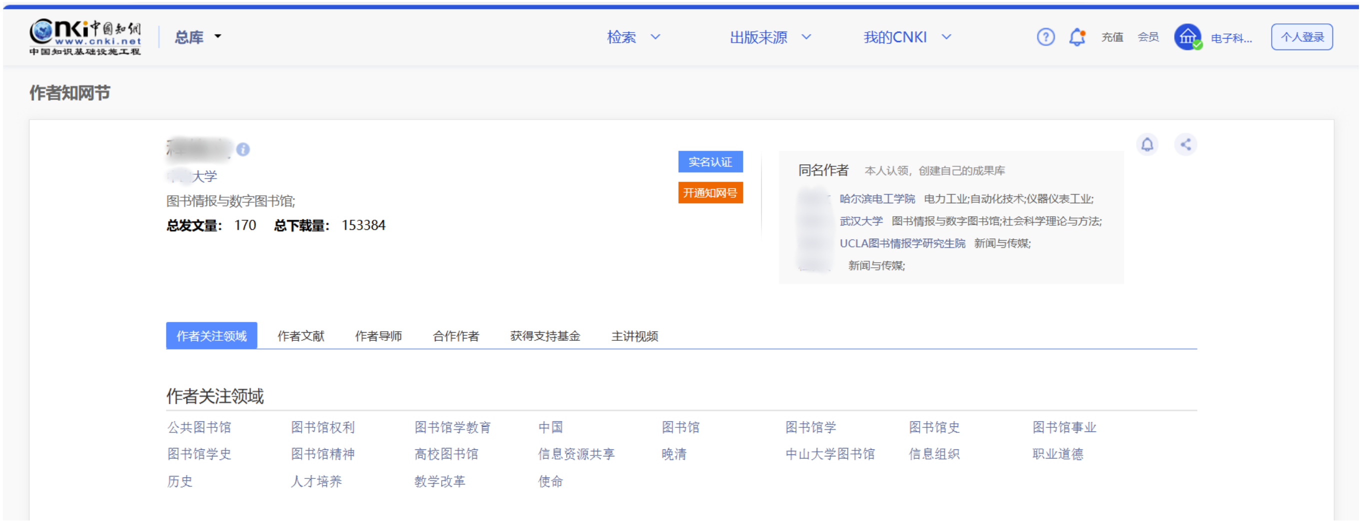
个人简介：[redacted]

工作经历：1986年7月分配到[redacted]大学信息管理学院任教
1992.10-1993.10 美国[redacted]大学洛杉矶分校（UCLA）图书馆学信息科学研究生院访问学者 1994-1996[redacted]大学信息管理系副主任
1996-1998[redacted]大学信息管理系主任
1998-2019[redacted]大学图书馆馆长
1997-1999[redacted]大学信息科学与技术学院副院长
1998.8-1999.1 美国[redacted]大学（UIUC）图书馆学信息科学研究生院访问学者
2000-2003[redacted]大学网络教育学院院长
2001.10-2002.1 美国[redacted]大学哈佛燕京图书馆和哈佛燕京学社访问学者
2003-2005[redacted]大学信息与网络中心主任
2005-2007[redacted]大学资讯管理系主任
2007-2010[redacted]大学传播与设计学院院长
2011.3 - 2012.3[redacted]大学总务处处长
2011.9-2012.11[redacted]大学校长助理
2015.6- 2017.9[redacted]大学校长助理 现任[redacted]大学信息管理学院教授，博士生导师 现任国家文化遗产与文化发展研究院院长"

Figure 6. Researcher 7's basic information window 1.

After comparing data sources, the basic information was found to be accurate. The school's official website confirmed the accuracy of researchers' photos, titles, fields, profiles, education, and work experience.

Comparing the basic information of Researcher 7 generated by our system (**Figure 6**) with his CNKI profile (**Figure 7**), we observe that the system's digital profile includes a photograph, enhancing the researcher's representation. The system also uses a word cloud to display the researcher's areas of focus, offering a more intuitive view. Additionally, it provides detailed information on the researcher's educational and professional background, allowing users to gain a comprehensive understanding of the researcher.



作者知网节

[redacted]
[redacted]大学
图书情报与数字图书馆;
总发文量: 170 **总下载量:** 153384

实名认证
开通知网号

同名作者 本人认领，创建自己的成果库
[redacted]
哈尔滨电工学院 电力工业;自动化技术;仪器仪表工业;
武汉大学 图书情报与数字图书馆;社会科学理论与方法;
UCLA图书情报学研究生院 新闻与传媒;
新闻与传媒;

作者关注领域 作者文献 作者导师 合作作者 获得支持基金 主讲视频

作者关注领域

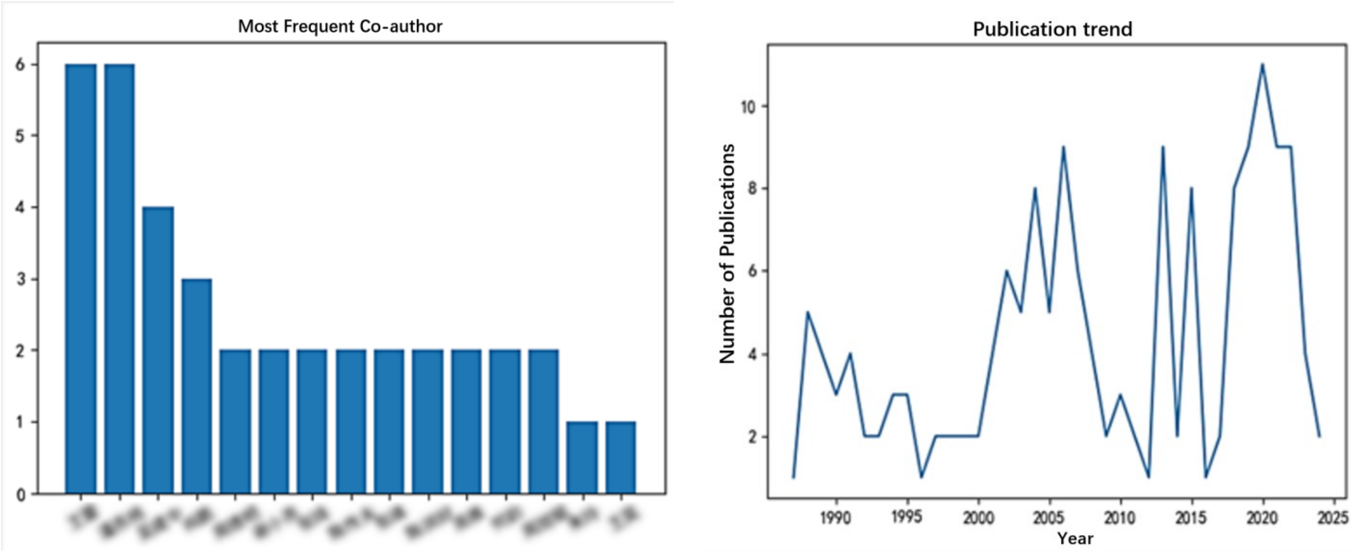
公共图书馆 图书馆权利 图书馆学教育 中国 图书馆 图书馆学 图书馆史 图书馆事业
图书馆学史 图书馆精神 高校图书馆 信息资源共享 晚清 中山大学图书馆 信息组织 职业道德
历史 人才培养 教学改革 使命

Figure 7. Researcher 7’s personal profile on CNKI.

Figure 8 shows window 2, which contains the researcher’s author collaborations, publication trends and citation trends.

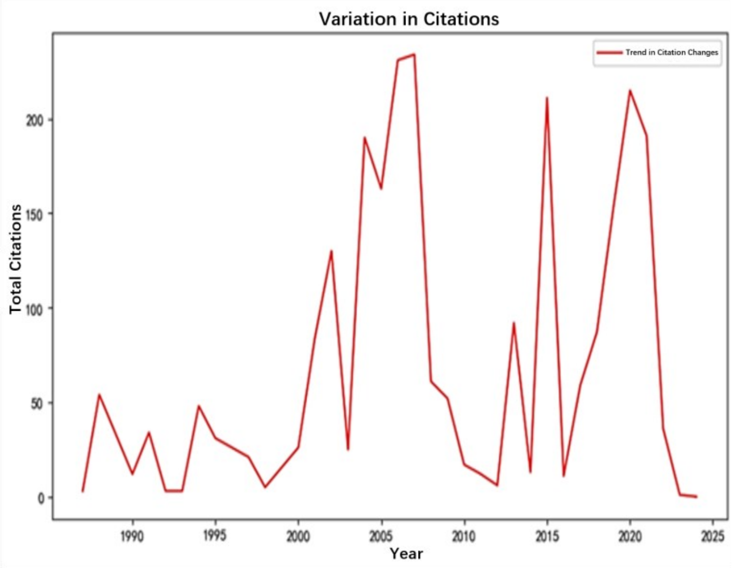
The comparison with Wanfang Data-Scholars’ Knowledge Pulse showed that the H-value, co-authors (**Figure 8a**), and publication trends (**Figure 8b**) are accurate. The citation trend (**Figure 8c)** is provided for reference only, due to differing statistical methods across websites.

The changes in Researcher 7’s publications and citations can be clearly seen in **Figure 8b,c**. Excluding the data at the current stage (2024), the trend of publication shows that the pattern of publication of Researcher 7 is an overall increase, which may mean that his/her research work is continuous and effective, and his/her research activities are more active. From the trend of citation changes, it can be seen that his/her citations are particularly high between 2005 and 2010 and around 2020, which indicates that his/her academic achievements in these time periods have been recognized and cited by many peer scholars, and have had a positive impact on the academic community.



(a)

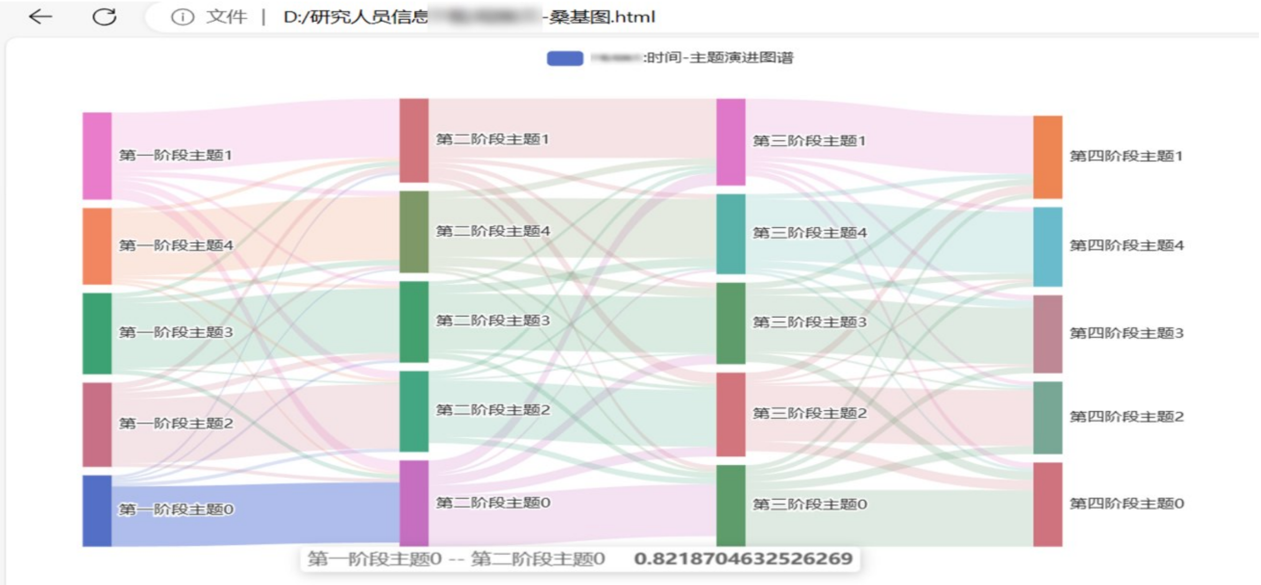
(b)



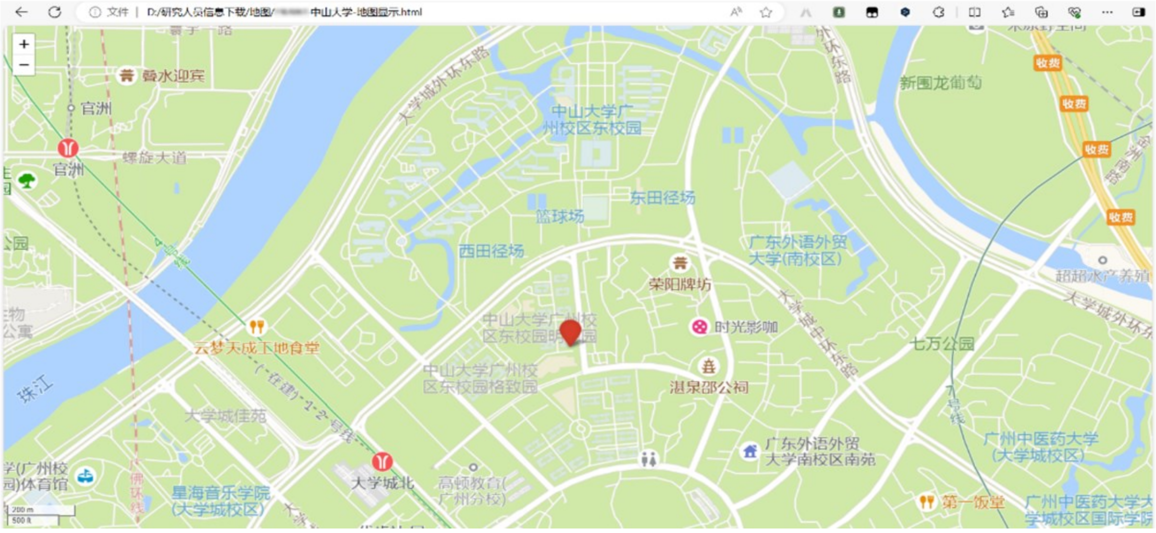
(c)

Figure 8. (a) Publication trend; **(b)** change trend of citations; **(c)** researcher 7’s co-author.

Figure 9a shows HTML page 1, which displays the time topic evolution of the researcher, and the similarity between stages can be viewed with the mouse to understand the time-topic evolution. **Figure 9b** shows HTML page 2, which calls the API of Baidu map through the input researchers’ work unit to calculate the latitude and longitude coordinates of the unit’s address and project them into the map.



(a)



(b)

Figure 9. (a) Time-topic evolution Sankey chart; **(b)** map of researcher's workplaces.

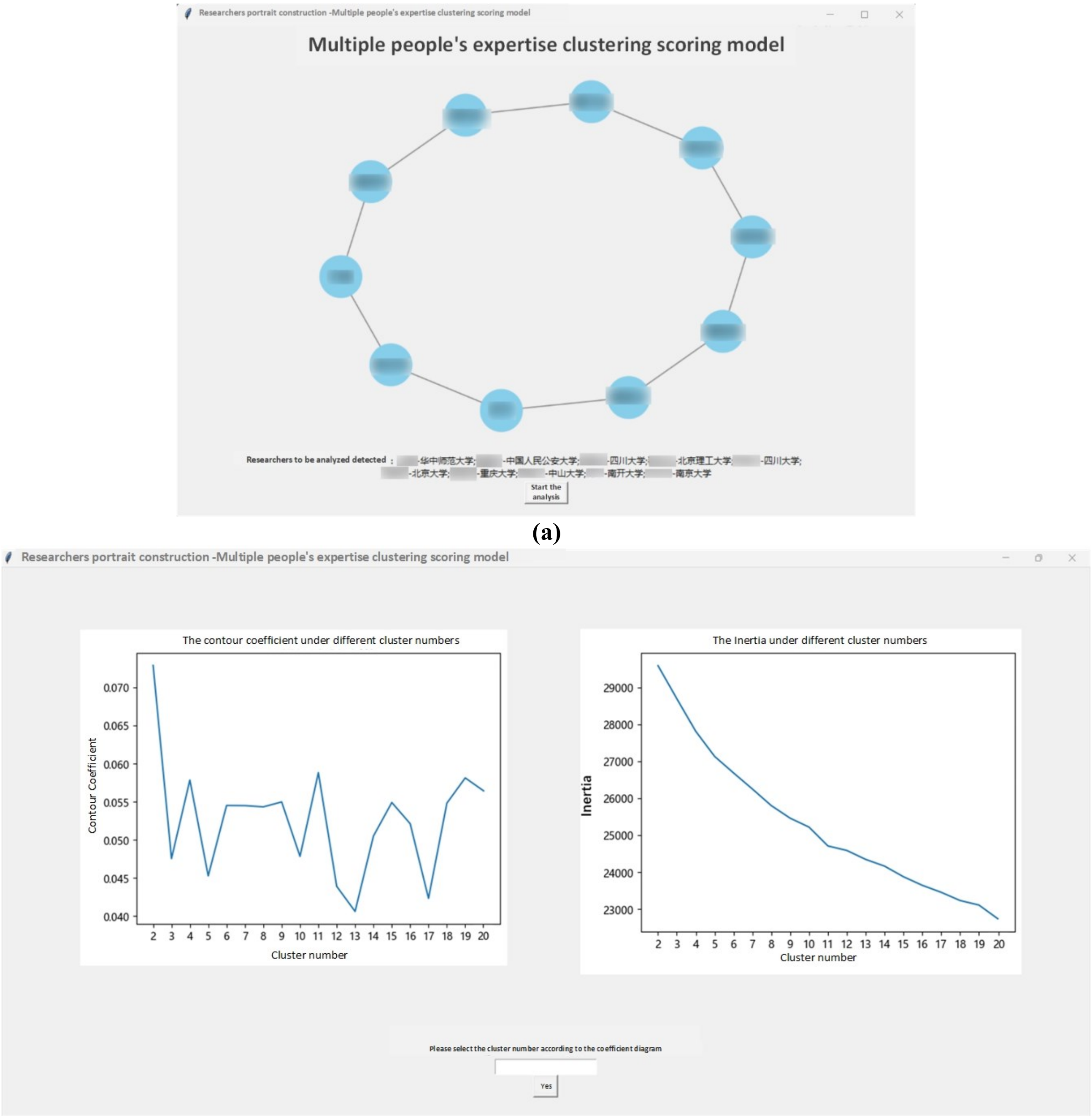
4.2. Domain clustering evaluation of multiple researchers

In the multiplayer mode, the system can cluster the expertise tag words of multiple researchers into domains and score them. In this section, we will use the 10 researchers selected in the experiments of this paper as an example for illustration.

Under the clustered scoring mode for multiple individuals' domains of expertise. the system reads the local data for display (**Figure 10a**), and the user can proceed to the next step after confirming that the reading is correct.

According to the method in 3.4., the system uses the K-means module of the scikit-learn library to cluster the expertise tag words of the 10 researchers. According to Equation (5), the inertia class of K-means is called to calculate the Inertia index, and according to Equation (6), the silhouette score function of K-means is called to calculate the profile coefficient. Inertia index and contour coefficients are calculated by the number of clusters from 2 to 20, and line plots are drawn. As shown in **Figure 10b**, the system outputs a line graph of the contour coefficients and a line graph of the Inertia index. A higher Silhouette Coefficient indicates that data points are more similar to their own cluster and less similar to other clusters. It is evident that the Silhouette Coefficient is high when the number of clusters is 4, 11, and 19. Additionally, the Inertia index line graph shows a significant change when the number of clusters is 11. Based on the 'elbow principle', the number of clusters selected in this paper is 11.

During the analysis process the system first calculates the Euclidean distance between each expertise tag word and the cluster center for each researcher using Equation (7). This distance is then converted to a similarity degree, which indicates the degree of similarity between the expertise tag word and the domain. Finally, the system selects the 10 words in the clustered domain that correspond to the expertise tag words closest to the cluster centers. **Table 2** lists the 10 expertise tag words within the 11 domains of the section that correspond to the closest match to the cluster cores.



Researchers portrait construction -Multiple people's expertise clustering scoring model

Multiple people's expertise clustering scoring model

Researchers to be analyzed detected ：[redacted]-华中师范大学;[redacted]-中国人民公安大学;[redacted]-四川大学;[redacted]-北京理工大学;[redacted]-四川大学;
[redacted]-北京大学;[redacted]-重庆大学;[redacted]-中山大学;[redacted]-南开大学;[redacted]-南京大学

(a)

Researchers portrait construction -Multiple people's expertise clustering scoring model

Please select the cluster number according to the coefficient diagram

(b)

Figure 10. (a) Analyze interface; **(b)** inertia index and silhouette coefficient line chart.

Table 2. Partial domain expertise tag words and domain overview.

Serial Number	Most Similar Expertise Tag Words	Domain Overview
0	Intelligence Studies; Intelligence; Intelligence Values; Intelligence Analysis Methods; Intelligence Analysis Models; Intelligence Collection; Intelligence Surveillance; Library and Intelligence Organizations; Intelligence Exchange; Library and Intelligence Career	Intelligence
1	Technology Applications; Machine Learning; Algorithmic Recommendations; Data Driven; Data Structured Analysis; Random Forest; Blockchain Technology; User Behavior Models	Computer technology and applications
2	Libraries; public libraries; college libraries; national libraries; library modernization; community libraries; book reservations; book donations; collection agencies; digital libraries	Library

* The table lists only some of the 10 domains and is not complete.

The value of amplification factor C is set to 15 to measure the expertise of a researcher in a particular research area. Once the system is done, the user can use the pyplot library to visualize the expertise of the researchers in each area. This window (**Figure 11**) displays the rankings of 10 researchers across 11 different clustering domains.

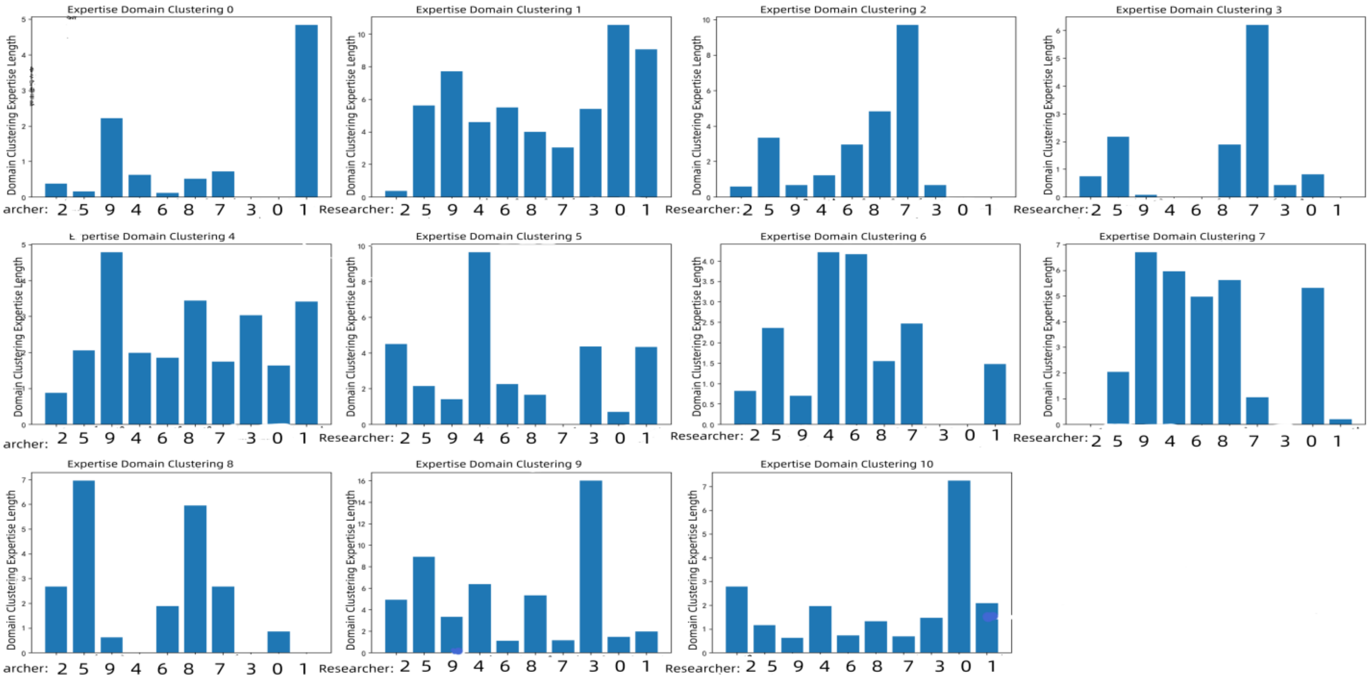


Figure 11. Clustering domain evaluation results.

By visualizing the evaluation of these 10 researchers in different domains through the bar chart, we can intuitively observe their focus in specific domains. This provides valuable data support for aligning research projects with appropriate team members and ensures the quality of research collaborations.

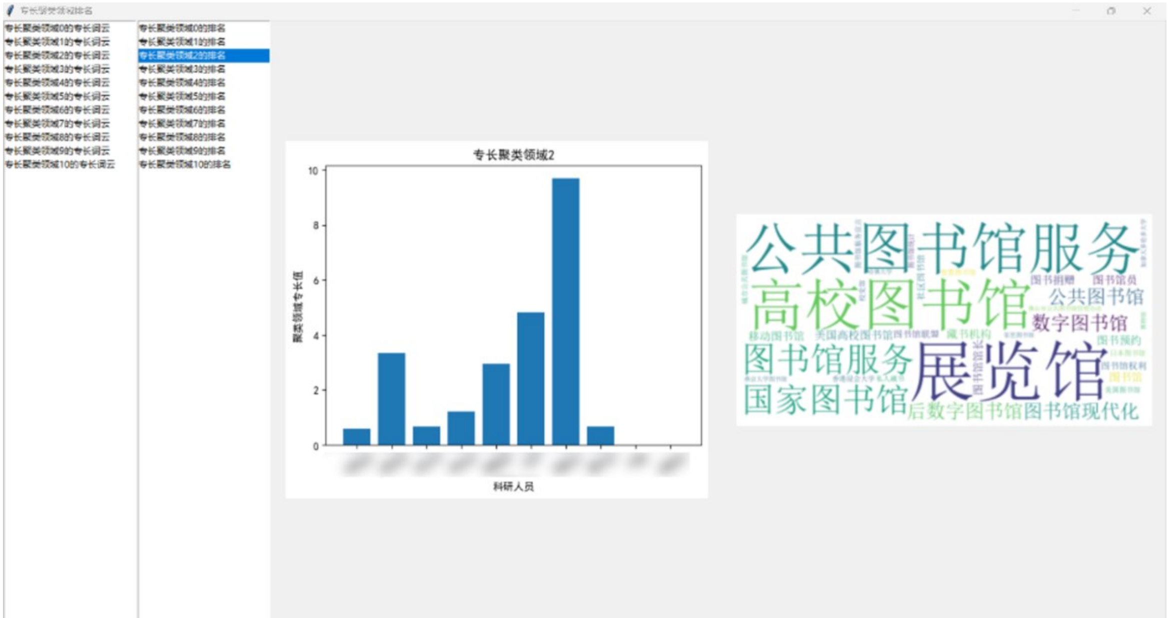


Figure 12. Expertise tag words clustering domains and scoring results presentation.

Taking the score results for Domain 2 in **Figure 12** as an example, the word cloud indicates that this domain is primarily related to 'exhibition halls.' It is evident that Researcher 6 has the highest score in this domain, demonstrating a clear advantage in this research domain.

5. Discussion

Figure 13 shows the research interest river diagram of Researcher 7 in the AMiner platform. Compared with the experimentally obtained Time-Topic Evolution Sankey Chart (**Figure 9a**), the topics described with topic words are more comprehensive and detailed, and the topic evolution in each time period is more clearly reflected. To a certain extent, the Time-Topic Evolution Sankey Chart can better reflect the research trends of researchers at each stage.

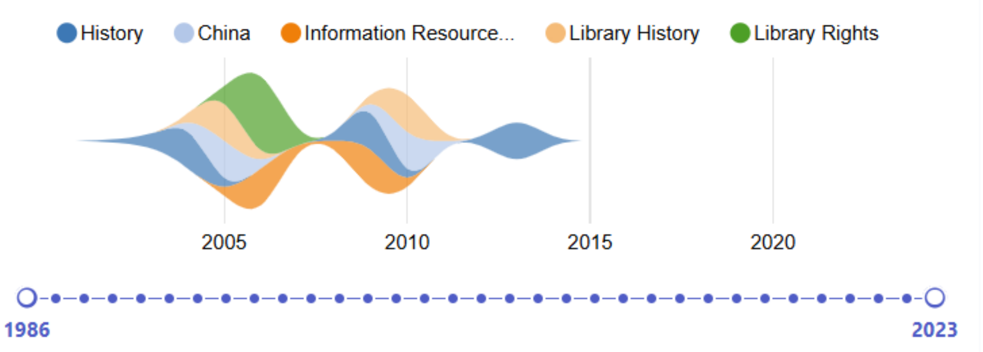


Figure 13. Research interest river of researchers 7.

To verify the results and explore additional patterns and information that can be reflected, we searched CNKI with the topic of "intelligence" and counted the information of academic papers published by 10 researchers according to the value of expertise in ascending order, as shown in **Table 3**.

Table 3. Information on researchers' papers related to "intelligence".

Researcher Number	Expertise Value	Number of Relevant Papers	Ratio of Related Papers to Total papers	Total Number of Citations	Highest Number of Citations Per Paper
Researcher 1	4.837894	41	0.732	753	198
Researcher 9	2.217244	57	0.141	678	340
Researcher 7	0.713881	21	0.132	197	94
Researcher 4	0.615767	12	0.132	405	157
Researcher 8	0.516278	47	0.131	609	174
Researcher 2	0.371626	4	0.068	239	157
Researcher 5	0.152488	9	0.082	123	67
Researcher 6	0.1182	9	0.079	98	50
Researcher 3	0	1	0.017	15	15
Researcher 0	0	0	0.000	0	0

In order to facilitate observation, the domain expertise value of Researcher 3 and Researcher 0 was set to 0.01. At the same time, the number of relevant papers of Researcher 0 was set to 0.1, and the data were processed using logarithms with a base of 10, so as to obtain the expertise value of each scientific researcher and the information of relevant papers in the domain 0 as shown in **Figure 14**.

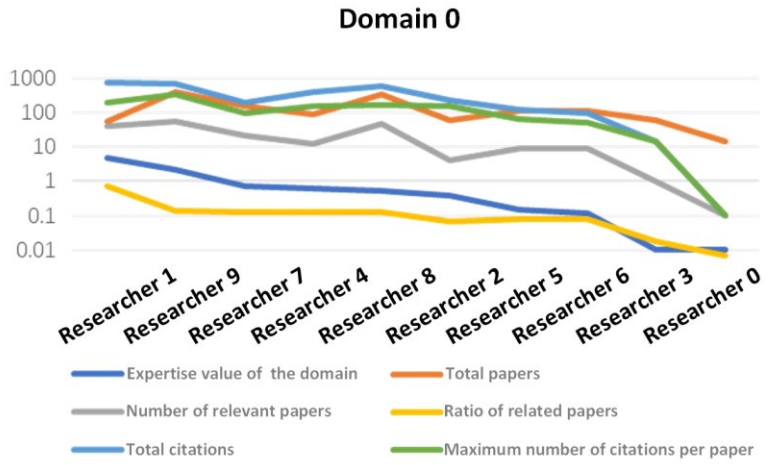


Figure 14. Domain 0 expertise value and relevant paper information.

Based on **Table 3** and **Figure 14**, it can be concluded that the value of researchers' expertise in the domain is roughly positively and linearly related to their ratio of relevant papers published in the domain to the total number of papers published. The strength of the linear relationship was measured using the Pearson correlation coefficient r:

$$r = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}} \tag{8}$$

where X_i is the value of domain expertise, and $\overline{X}$ is the average of the domain expertise values; Y_i is the ratio of relevant papers published in the domain to the total number of papers published, and $\overline{Y}$ is its average; and n is the number of domains.

In statistics, Confidence Interval (CI) is used to estimate the range of an unknown parameter. Based on sample data, a confidence interval infers with some probability (usually 95% or 99%) the likelihood that the overall parameter falls within a certain range. In short, confidence intervals provide us with a range within which we can be confident that the true value of the overall parameter lies.

Based on the formula for the correlation coefficient r, we calculated the r values for the 11 domains clustered in the current experiment and determined the confidence intervals for these 11 domains to range from 0.832 to 0.917 at the 95% confidence level. The mean r value is 0.874, which falls within that confidence interval. This further demonstrates that the current sample mean is a good representation of the overall mean, validates the sensitivity of this scoring method to the proportion of relevant papers to the total number of published papers, and to some extent reflects the research focus of scientific researchers in a particular field.

To validate the applicability of this method, this paper selected 10 researchers with backgrounds in science and engineering, as detailed in **Table 4**, for experimental verification. After calculating the expertise tag words, the data were categorized into 8 distinct domains based on the metrics of Inertia and Contour coefficient. **Table 5** lists the 10 closest words to the cluster centers for each of the expertise tag words within the 8 domains of the section that correspond to the closest match to the cluster cores.

Table 4. 10 researchers with backgrounds in science and engineering.

Researcher Number	Work Unit	Acquisition papers
Researcher 10	Xiangtan University	25
Researcher 11	Shandong University of Science and Technology	68
Researcher 12	Hebei University of Technology	71
Researcher 13	Southwest Jiaotong University	95
Researcher 14	Beijing University of Posts and Telecommunications	108
Researcher 15	Huazhong University of Science and Technology	132
Researcher 16	Shanxi University	151
Researcher 17	Sichuan University	238
Researcher 18	Central South University	252
Researcher 19	Wuhan University of Technology	387

Table 5. Partial domain expertise tag words and domain overview.

Serial Number	Most Similar Expertise Tag Words	Domain Overview
0	Neural Networks; Convolutional Neural Networks; Backpropagation Networks; Graph Neural Network; Recurrent Neural Networks; Long Short-Term Memory; Deep Belief Network; Feedforward Neural Networks; Deep Learning; Attention Mechanism	Artificial Intelligence
1	Information Security; Security Policy; Computer Security; Cloud Security; Service Attack; Intrusion Detection System; IoT Security; Security Audit; Malware; Virus protection; Ransomware	Cybersecurity
2	Sensor; Filter; Signal-to-Noise Ratio; Bit Rate; Generator; Load Balancing; High Precision; Recognition Rate; Offset; Modular	Sensor Technology

* The table lists only some of the 10 domains and is not complete.

Table 6. Results of Pearson correlation coefficient r calculation for 8 domains

Number	Domain	Calculated Value r
0	Artificial Intelligence	0.873
1	Cybersecurity	0.797
2	Sensor Technology	0.777
3	Signal Processing	0.595
4	Data Management	0.641
5	Image Processing	0.577
6	Control Engineering	0.694
7	Software Engineering	0.696
Average Calculated Value		0.706

The Pearson correlation coefficient r was calculated for the domain expertise values of researchers with science and engineering backgrounds across 8 domains, as shown in **Table 6**, in relation to the ratio of relevant papers published to the total number of papers**.** The 95% confidence interval for these 8 samples was determined to be between 0.611 and 0.801. The mean r value across the 8 domains was 0.706, which falls within the 95% confidence interval. This outcome indicates that the method reflects a good degree of focus among researchers with a Science, Technology, Engineering, and Mathematics (STEM) background, but it is relatively lower compared to researchers with a social science background. Upon inspection, it was found that due to the abundance of specialized vocabulary and abbreviations in the field of science and engineering, the classification effect of these terms by the pre-trained BERT model is not as effective as that of more semantically clear professional background terms. However, overall, it is generally acceptable. Therefore, this demonstrates that the scoring method is sensitive to the proportion of relevant papers in the total number of published papers and can, to a certain extent, reflect the degree of focus of researchers in a particular field.

Therefore, we can consider that using this method to describe and score domain specific knowledge can to some extent reflect the researchers' level of focus on a certain field, and provide some support for expert discovery and academic evaluation, laying the foundation for further in-depth research.

6. Conclusion

In this paper, we designed and implemented an interactive digital profile system of Chinese researchers based on unsupervised machine learning using the Python programming language. The system constructs a multi-dimensional profile of researchers by analyzing their basic and behavior information, and collects data in real time from the CNKI and AMiner platforms by combining Selenium and web crawler technology. Subsequently, the system employs the TF-IDF algorithm, BERT pre-training model, DTM dynamic model, and K-means clustering algorithm to quantitatively evaluate researchers' research dynamics and domain expertise. The experimental results indicate that the system can effectively construct digital profiles of researchers in both the Chinese social sciences and STEM fields. The unsupervised learning approach performs well within this system, ensuring high computational efficiency. Additionally, the interactive system provides robust data support for academic evaluation and talent management.

However, the current study has certain limitations. Firstly, there is a time lag from the submission of a paper to its final publication, so the system has an unavoidable delay effect. In addition, the current system mainly focuses on constructing profiles of researchers' basic information and behavior information, and is limited to Chinese subjects, which lacks language breadth. Future research will aim to extend the scope of data collection to cover areas such as social networks and conference participation, and optimize real-time data updating and processing methods to cope with the existing delay issues. To enhance the system's global adaptability, we plan to extend it to international researchers and add multilingual support. Furthermore, deep learning techniques and additional weighting factors will be introduced in future studies to further enhance the accuracy of domain clustering and expertise evaluation, thereby enabling a comprehensive understanding and analysis of researchers' expertise and research dynamics. Moreover, the pre-trained BERT model tends to perform relatively poorly with certain abbreviations of technical terms. Future research could explore training a specialized BERT model, using a combination of unsupervised and supervised methods to further improve accuracy.

We expect the system to play a greater role in a wider range of research scenarios and provide more accurate data support for academic research and collaboration worldwide.

Author contributions: Conceptualization, PX, YY and TH; methodology, TH and YY; software, PX and YY; validation, PX and SL; formal analysis, PX and YY; investigation, YY and SL; resources, YY and SL; data curation, YY and SL; writing—original draft preparation, PX; writing—review and editing, PX and SL; visualization, PX and YY; supervision, LY and JZ, project administration, LY and JZ; funding acquisition, TH and LY. All authors have read and agreed to the published version of the manuscript.

Conflict of interest: The authors declare no conflict of interest.

References

Al-Shamri, M. Y. H. (2016). User profiling approaches for demographic recommender systems. Knowledge-Based Systems, 100, 175–187. https://doi.org/10.1016/j.knosys.2016.03.006

Blei, D. M., & Lafferty, J. D. (2007). Correction: A correlated topic model of Science. The Annals of Applied Statistics, 1(2). https://doi.org/10.1214/07-aoas136

Boussaadi, S., Aliane, D. H., & Abdeldjalil, P. O. (2020). The Researchers Profile with Topic Modeling. In: Proceedings of 2020 IEEE 2nd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS); 2–3 December 2020; Kenitra, Morocco.

Bulut, Z. A., & Doğan, O. (2017). The ABCD typology: Profile and motivations of Turkish social network sites users. Computers in Human Behavior, 67, 73–83. https://doi.org/10.1016/j.chb.2016.10.021

Chamorro-Padial, J., & Rodríguez-Sánchez, R. (2023). The relevance of title, abstract, and keywords for scientific paper quality and potential impact. Multimedia Tools and Applications, 82(15), 23075–23090. https://doi.org/10.1007/s11042-023-14451-9

Chavez, J. V., Libre, J. M., Gregorio, M. W., et al. (2023). Human resource profiling for post-pandemic curriculum reconfiguration in higher education. Journal of Infrastructure, Policy and Development, 7(2), 1975. https://doi.org/10.24294/jipd.v7i2.1975

Chen, L., Guo, S., Teng, G., et al. (2019). Research on the focus and migration of researchers' study topics (Chinese). Journal of the China Society for Scientific and Technical Information, 12, 9-17. https://doi.org/10.3772/j.issn.1673-2286.2019.12.002

de Campos, L. M., Fernández-Luna, J. M., Huete, J. F., et al. (2020). Automatic construction of multi-faceted user profiles using text clustering and its application to expert recommendation and filtering problems. Knowledge-Based Systems, 190, 105337. https://doi.org/10.1016/j.knosys.2019.105337

Gao, J., Peng, B. (2021). Research on knowledge discovery methods for cultural relics information resources based on topic identification (Chinese). Information Science, 39(4), 9-14.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences, 102(46), 16569–16572. https://doi.org/10.1073/pnas.0507655102

Holanda, O., Elias, E., Costa, E., et al. (2013). Towards an Agent-Based Approach for Automatic Generation of Researcher Profiles Using Multiple Data Sources. In: Proceedings of 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT). pp.163–166.

Jia, T., Wang, D., & Szymanski, B. K. (2017). Quantifying patterns of research-interest evolution. Nature Human Behaviour, 1(4). https://doi.org/10.1038/s41562-017-0078

Lei, W. (2017). Enhanced text feature representation for short text topic modeling (Chinese) [Master's thesis]. Nanjing University of Aeronautics and Astronautics.

Li, S. (2023). Research on user profiling technology for university researchers (Chinese). In: Proceedings of the 27th Annual Conference on New Network Technologies and Applications of the China Computer Users Association Network Application Branch; November 2023; Zhenjiang, China. pp. 498-501.

Liu, C. (2018). Research on an internal threat detection framework based on user profiling technology (Chinese) [Master's thesis]. Information Engineering University.

Noureddine, H., Jarkass, I., Hazimeh, H., et al. (2015). CARP: Correlation Based Approach for Researcher Profiling. In: Proceedings of the 27th International Conference on Software Engineering and Knowledge Engineering; 6–8 July 2015; Pittsburgh, USA.

O'Leary, D. E. (2021). An Analysis of Information Systems Researcher and Collaboration Rankings. Journal of Organizational Computing and Electronic Commerce, 31(3), 270–292. https://doi.org/10.1080/10919392.2021.1975477

Özçift, A., Akarsu, K., Yumuk, F., et al. (2021). Advancing natural language processing (NLP) applications of morphologically rich languages with bidirectional encoder representations from transformers (BERT): an empirical case study for Turkish. Automatika, 62(2), 226–238. https://doi.org/10.1080/00051144.2021.1922150

Papaevangelou, O., Syndoukas, D., Kalogiannidis, S., et al. (2023). Efficacy of embedding IT in human resources (HR) practices in education management. Journal of Infrastructure, Policy and Development, 8(1). https://doi.org/10.24294/jipd.v8i1.2371

Pottier, P., Lagisz, M., Burke, S., et al. (2024). Title, abstract and keywords: a practical guide to maximize the visibility and impact of academic papers. Proceedings B, 291(2027), 20241222. https://doi.org/10.1098/rspb.2024.1222

Sateli, B., Löffler, F., König-Ries, B., et al. (2017). ScholarLens: extracting competences from research publications for the automatic generation of semantic user profiles. PeerJ Computer Science, 3, e121. Portico. https://doi.org/10.7717/peerj-cs.121

Song, P., Long, C., Ni, X., et al. (2022). Research on the method of identifying academic expertise of researchers based on the iceberg model (Chinese). Data Analysis and Knowledge Discovery, 50-60.

Tang, J. (2016). AMiner: Toward understanding big scholar data. In: Proceedings of the ninth ACM international conference on web search and data mining. pp. 467-467.

Tang, J., Li, J., Zhang, K., et al. (2018). AMiner: A big data mining and service platform for scientific and technological information (Chinese). China Science and Technology Achievements, 19(13), 57-58. https://doi.org/10.3772/j.issn.1009-5659.2018.13.026

Wang, Q. (2019). Information push and mining model based on time-varying neighborhood system (Chinese) [Master's thesis]. Southwest Jiaotong University.

Xu, L., Zhang, J., Zhang, C., et al. (2024). Beyond extraction accuracy: addressing the quality of geographical named entity through advanced recognition and correction models using a modified BERT framework. Geo-Spatial Information Science, 1–19. https://doi.org/10.1080/10095020.2024.2354229

Yimam-Seid, D., Kobsa, A. (2003). Expert-Finding Systems for Organizations: Problem and Domain Analysis and the DEMOIR Approach. Journal of Organizational Computing and Electronic Commerce, 13(1), 1-24. https://doi.org/10.1207/S15327744JOCE1301_1

Zhang, Y., Huang, J., Wang, G. (2019). A method for constructing a three-dimensional accurate profile of researchers' scientific behavior considering global and local information (Chinese). Journal of Information Science, 38(10), 1012-1021.

Zhao, H., Hua, B., He, H. (2020). Science and Technology Intelligence User Profile Label Generation and Recommendation (Chinese). Journal of the China Society for Scientific and Technical Information, 39(11), 1214-1222.

Zhao, Y. (2009). Research on K-means clustering mining method based on genetic algorithm (Chinese) [Master's thesis]. Qingdao University of Science and Technology.

Zhu, L., Ma, B., Zhao, X. (2010). Cluster validity analysis based on silhouette coefficient. Computer Applications, 30(12), 139-141.

Zou, L., He, Z., Zhou, C., et al. (2024). Multi-class multi-label classification of social media texts for typhoon damage assessment: a two-stage model fully integrating the outputs of the hidden layers of BERT. International Journal of Digital Earth, 17(1). https://doi.org/10.1080/17538947.2024.2348668