Postprint: Research on the Construction Method of Researcher Profiles Integrating Multi-source Data

Authors: Fan Xiaoyu, Dou Yongxiang, Zhao Pengwei, Zhou Xiao

Date: 2023-08-27T00:00:00+00:00

Abstract

[Purpose/Significance] In the era of big data, there is a need for the datafication of individuals, and researchers also require datafication. The establishment of researcher profiles plays a significant role in enabling research management to comprehensively understand researcher information and objectively evaluate their research competence. It can also serve as a foundation for analyzing researcher behavior or expert recommendation, thus improving research management efficiency. [Method/Process] First, the concept of researcher profile is proposed, considering it as a collection of tags that describe researcher information. Second, based on data from multiple heterogeneous sources such as personal homepages, CNKI, and funding databases, a method for constructing researcher profiles by fusing multi-source data is proposed, which formally describes researcher information from three aspects: basic attributes, research preferences, and research relationships, extracts tags from each dimension, and visualizes the profile. Finally, the feasibility of the researcher profile construction method is demonstrated by taking two researchers from domestic and international contexts as examples. [Result/Conclusion] The construction of researcher profiles is applicable to researchers both domestically and internationally, capable of comprehensively describing researcher information and presenting it intuitively.

Full Text

Preamble

ChinaXiv Cooperative Journal, Vol. 62, No. 15, August 2018

Research on the Construction Method of Researcher Profiles Integrating Multi-source Data

Fan Xiaoyu, Dou Yongxiang, Zhao Pengwei, Zhou Xiao
School of Economics and Management, Xidian University, Xi'an 710071

Abstract

[**Purpose/Significance**] In the era of big data, people need to be digitized, and researchers are no exception. The establishment of researcher profiles is crucial for research management to comprehensively understand researcher information and objectively evaluate their research levels. It can serve as a foundation for analyzing researcher behavior or expert recommendation, thereby improving the efficiency of research management. [**Method/Process**] This paper first proposes the concept of researcher profiles as a collection of labels describing researcher information. Based on data from multiple heterogeneous sources including personal homepages, CNKI, and funding networks, we propose a method for constructing researcher profiles that integrates multi-source data. We formally describe researcher information from three dimensions: basic attributes, research preferences, and research relationships, extract labels for each dimension, and visualize the profiles. Finally, we demonstrate the feasibility of this method using two researchers, one domestic and one international, as examples. [**Result/Conclusion**] The construction of researcher profiles is applicable to both domestic and international researchers, enabling comprehensive description and intuitive visualization of researcher information.

Classification Number: G203
Keywords: researcher profile, multi-source data, user model

Researchers are the main actors in scientific activities, and their information constitutes an important knowledge resource that plays a pivotal role in scientific research, project evaluation, achievement transformation, and decision-making consultation. The *Analysis of China's Scientific and Technological Human Resources Development in 2015* released in June 2017 indicates that by 2015, China's scientific and technological human resources continued to increase, reaching a total of 79.15 million, maintaining China's position as the world's largest country in scientific and technological human resources. However, with the growing number of researchers, corresponding problems have emerged. On the one hand, although China has established multiple researcher information databases, various types of information are stored separately by different departments, resulting in fragmented data and a lack of integration and correlation, which leads to low utilization of information. Research management departments must search through various databases to understand a researcher, making it impossible to gain an intuitive and rapid understanding. On the other hand, in research evaluation, the phenomenon of "paper-only" indicators or overemphasis on paper metrics is widespread, with identical evaluation standards applied to different types of researchers, ignoring contributions in other areas. The big data era requires the digitization of "people," and researchers are no exception. The rapid development of internet and information technology has made researcher interaction and communication increasingly convenient, while

researcher-related information has become dynamic, massive, and multi-source heterogeneous. Effective collection, management, and analysis of researcher information will help grasp the current state of research, uncover key influencing factors in research projects, build a "multiplier" for modern scientific and technological innovation and a "think tank" for scientific decision-making, and fundamentally change traditional models of scientific and technological work management and decision-making. Therefore, this paper proposes the concept of researcher profiles for the big data of scientific and technological personnel information, which can fully reveal individual characteristics and preferences of researchers, accurately portray "thousands of faces for thousands of people," and provide personalized services and precise recommendations, as well as authentic and effective reference basis for decision-makers.

1 Research Status

In recent years, an increasing number of scholars have recognized the importance of researcher information and how to establish a unified framework for describing researcher information. Driven by research on the semantic web and linked data, some scholars have extracted person attributes and relationships from large-scale personnel data to construct person ontologies, modeling personnel information and interpersonal relationships to form a mechanism for sharing and semantic description. In the field of science and technology management, Europe adopted the EuroCRIS system to build a unified description model CERIF, integrating scientific research projects, experts, achievements, institutions, instruments, and other scientific and technological resources for interconnection and interoperability. C. Moreira et al. used D-S evidence theory and Shannon entropy to obtain expert information from expert achievements, citation networks, and basic profiles. L. Yang et al. described researcher information from five aspects: projects participated in, awards, published articles, monographs, and granted patents, and built a semantic and knowledge reasoning-based expert system. T. H. Trong et al. used ODP (Open Directory Project) to establish a scholar semantic search space and generate information ontologies for describing researchers.

Domestically, Li Gang et al. designed an expert feature identification method based on multi-source information fusion, integrating knowledge, web, and social network sensors. Wang Yuefen et al. introduced advanced expert retrieval technology, social network analysis technology, and visualization technology into expert database construction, describing researcher information from basic characteristics and relationships, and designed a construction process for science and technology consulting expert databases. Song Peiyan et al. built a scientific expert semantic model based on knowledge organization theory, used RDF for formal description and empirical research, and finally generated an expert information database with strong standardization and semantic relationships. Lu Wei et al. obtained expert personal and relationship information through network databases, search engines, and expert recommendation forms, complet-

ing expert organization work and system construction respectively. J. Tang et al. designed the AMiner system, which automatically obtains researcher-related information from massive literature and internet information to establish researcher description pages, providing diverse services such as search, academic evaluation, collaborator recommendation, reviewer recommendation, and topic trend analysis.

Overall, both domestically and internationally, significant achievements have been made in describing researcher information. However, we find that existing studies have limitations: First, the description of expert information is limited to academic resources with a single perspective, lacking in-depth mining of researcher interests or research relationships, resulting in incomplete descriptive information. Second, although some studies provide comprehensive descriptions of expert information, research managers must examine the database item by item to understand researcher information, preventing a rapid and intuitive understanding. To address these issues, this paper proposes the concept of researcher profiles and its construction method.

2 Researcher Profile Based on Multi-source Dynamic Data

Multi-source data refers to data generated by different users and from different sources, with multiple presentation forms, describing the same theme. Researcher data often comes from multiple sources and is presented in different forms and perspectives, comprehensively describing researchers with complementary information from different sources. Some data are static while others are dynamic. Personal information is relatively stable and belongs to static information, whereas achievement information and research interests change due to environmental or demand influences, making them dynamic information.

The concept of user profiles was first proposed by A. Cooper, who believed that "user profiles can represent a real user and are user models built using real user data." This paper defines researcher profiles as a labeled, formalized user model abstracted from researchers' social attributes, research habits, and research behaviors. Following the construction process of user profiles, this paper proposes a researcher profile construction process, as shown in Figure 1 [Figure 1: see original paper].

The construction of researcher profiles is a dynamic update process achieved by regularly collecting various types of researcher information. The process is as follows: First, collect researcher information from multiple data sources, including demographic attribute data, research achievement data, and research preference data. Through data preprocessing, generate usable data for profile construction and store it in a researcher information database. Then, formally represent researcher information in vector form to build the researcher model. Finally, label each dimension of researcher information according to the model, update the researcher information labels, and present the researcher profile using visualization tools.

2.1 Data Collection

Data collected for Chinese researchers mainly includes demographic attribute data, research achievement data, research behavior preference data, research collaboration data, and research social data.

Demographic attribute data refers to the demographic statistical characteristics of researchers, including name, gender, date of birth, contact information, etc. Research achievement data is an indispensable part of researcher information, referring to academically significant outputs produced during scientific research, including journal papers, conference papers, academic monographs, patents, conference reports, etc. Research behavior preference data represents users' interest in certain research topics, while research collaboration data represents data generated through collaboration with others; both are obtained through analysis of research achievement information. Research social data is data generated by researchers in academic social networks. The specific content and collection methods for each type of data are shown in Table 1 .

2.2 Data Preprocessing

Data obtained from various sources cannot be directly used for profile construction and requires preprocessing. Table 2 lists the characteristics and existing problems of raw data of various types. Through preprocessing, these are transformed into data usable for constructing researcher profiles.

Unstructured data mainly exists in demographic attribute data. Researchers' basic information is often described in text form. To facilitate storage and profile construction, text data must first be converted into structured data, requiring named entity recognition in the text. Named entities mainly include name entities (organization names, person names, place names), time expressions (dates, times), and numerical expressions (monetary values, percentages). Among these, the recognition of organization names, person names, and place names is particularly challenging due to their open and evolving nature. Currently, multiple open-source Chinese language processing tools are available for direct use in named entity recognition, such as fudanNLP developed by Fudan University and the NLPIR segmentation system from the Chinese Academy of Sciences, both of which can be implemented through Java calls. The LTP system from Harbin Institute of Technology also provides a Python interface, allowing direct use of the pyltp module for named entity recognition to extract researcher information from text.

Structured data has problems such as missing data, duplicate data, and homonymy. Missing data can be supplemented through search engines like Baidu or by consulting the individuals themselves. Duplicate data requires deletion of redundant information to ensure uniqueness, with remaining information complementing each other. Homonymy issues require name disambiguation. Commonly used methods for name disambiguation include clustering-based disambiguation and entity linking-based disambiguation, which have been

extensively studied and are relatively mature.

Data integration involves consolidating data from multiple sources into a unified database. In raw data, different sources adopt different metadata standards, resulting in a lack of unified description of researcher information. The same attribute may use different field names in different databases. We use unified fields to describe researcher information, as shown in Table 3 .

Based on researcher data types, we use four databases to store this data: a basic attribute database, a research achievement information database, a research preference database, and a research relationship database. These components are interrelated. The basic attribute database stores basic demographic statistical attribute data of researchers, the research achievement information database stores various achievement information, the research preference database stores researchers' interest topics, and the research relationship database stores collaboration and social relationship data with researchers.

2.3 Researcher Profile Model Construction

Currently, studies describing researcher information have two main problems: (1) Single-dimensional description, focusing only on academic achievements while ignoring other aspects; (2) Lack of intuitive presentation of researcher information. To address these issues, this paper proposes a researcher profile model integrating multi-source data and instantiates the model, storing labels of each dimension in corresponding label libraries.

2.3.1 Researcher Profile Model A researcher profile is a multi-dimensional, multi-level user model. Based on data types in the researcher information database, this paper defines a triple as the vector space representation of user information:

$$\text{User} = \langle \text{Demographics}, \text{Interests}, \text{Relation} \rangle$$

where Demographics represents the basic attribute dimension, Interests represents the research preference dimension, and Relation represents the research relationship dimension. The multi-level researcher profile model is shown in Figure 2 [Figure 2: see original paper].

2.3.2 Tag Extraction and Weight Calculation **(1) Basic Model Tag Extraction.** In Section 2.3.1, we use $\text{Demographics} = \langle \text{BaseInfo}, \text{Edu}, \text{Org}, \text{Message}, \text{Achv} \rangle$ to represent the basic attribute model, consisting of demographic and research achievement dimensions. Since basic information in the research information database is concise, we can directly use database information as tags. Research achievement tags use achievement titles.

(2) Research Preference Tag Extraction. The research preference vector model is:

$$\text{Interests} = \{(\text{Topic}_1, t_1), (\text{Topic}_2, t_2), (\text{Topic}_3, t_3), \ldots, (\text{Topic}_n, R_n)\}$$

where Topic_n represents the nth interest topic of the researcher, and t_n represents the interest degree of the user for the nth topic. A larger t_n indicates higher interest. Since keywords are highly condensed summaries of article content, they can serve as primary representations of research topics. This paper uses keywords from researchers' published literature as Topic tags and calculates their weights sequentially. The weight calculation is as follows:

Since researchers' research topics are not static and change with the surrounding environment or subjective interests, the weight calculation for research topic tags combines preference weight and decay weight. Preference weight refers to the proportion of the tag among all tags, denoted by ω_i, where n_i represents the number of tag occurrences and N represents the total number of tags. For decay weight, we adopt the adaptive exponential decay function proposed by Y. Cheng et al. to represent the decay of researcher interest in a tag:

$$\theta_i = e^{-\ln 2 \times \frac{t-est}{hl}}$$

where t is the current time, est is the earliest time the tag appeared, and hl is the half-life of interest topic decay. The shorter the researcher's behavioral cycle for a topic (i.e., the research cycle for this tag), the smaller the hl, and the faster the interest decays; otherwise, the slower the decay. Therefore, recently studied topics are assigned greater weight, while older interests are assigned smaller weight.

Combining preference weight and decay weight, the comprehensive weight of a tag is:

$$t = \lambda\omega_i + (1-\lambda)\theta_i$$

where λ is a harmonic factor used to adjust the proportion of preference weight and decay weight. This approach captures both the user's preference degree for research topics and the time factor, reflecting the drift of researcher interests.

(3) Research Relationship Tag Extraction. In the research relationship vector model $\text{Relation} = \langle \text{ReTag}, R_{ui}, R_u \rangle$, ReTag is the node label in research relationships, R_{ui} is the relationship weight between user u and user i (represented by line thickness in the research relationship graph), and R_u is the contribution size of user u in the relationship (represented by node size in the research relationship graph).

ReTag node labels directly use the names of co-authors. We obtain all author names from researchers' published articles, deduplicate them, and use them as ReTag in the research relationship vector model.

When calculating R_u, we compute the contribution value based on the author's signature position in articles. This paper refers to the author contribution rate level allocation method, calculating each author's contribution weight according to author order, and finally superimposing the author's weight in each article to obtain the node's total weight. The level allocation method means that in co-authored literature, each author's weight decreases sequentially from first to last. Assuming an article has 5 co-authors, the contribution degrees from first to fifth are 5/15, 4/15, 3/15, 2/15, and 1/15 respectively. Therefore, for an article k with n co-authors, the contribution degree of the author ranked i is:

$$\omega_{ki} = \frac{n-i+1}{\sum_{i=1}^{n}(n-i+1)}$$

If the author has published m articles, the author's total contribution degree is:

$$R_u = \sum_{k=1}^{m} \omega_{ki}$$

Researcher data in this paper comes from multiple sources. To reflect the importance of different sources, we set contribution weights for authors from different sources. The weighted formula is:

$$R_u = \alpha_1 W_1 + \alpha_2 W_2 + ... + \alpha_p W_p$$

where α_p represents the proportion of literature from different sources to the total literature. Assuming researcher data comes from p sources, with $x_1, x_2, ..., x_p$ literature items from each source, then $x_1 + x_2 + ... + x_p$ equals the total.

R_{ui} represents the relationship strength between user u and user i, including both co-authorship and social relationships, denoted as CoR_{ui} and SoR_{ui} respectively, with $R_{ui} = 0.5 \times \mathrm{CoR}_{ui} + 0.5 \times \mathrm{SoR}_{ui}$. Co-authorship is represented by the proportion of co-authored articles to the total number of articles published by user u:

$$\mathrm{CoR}_{ui} = \frac{\mathrm{Paper}_{ui}}{\mathrm{Paper}_{u}}$$

where Paper_{ui} represents the number of co-authored papers between users u and i, and Paper_{u} represents the total number of papers published by user u. Social relationships are represented by a Boolean value: if user i is a friend of user u, then $\mathrm{SoR}_{ui} = 1$, otherwise 0.

2.4 Researcher Profile Update

Researcher profile updating involves updating basic information, research interest tags, and research relationship tags. Changes in basic information mainly include workplace, contact information, correspondence address, title, and position. Basic information updates can be performed by database managers periodically sending emails to researchers to remind them to update their information. When data in the researcher basic information database changes, corresponding tags also change accordingly.

Research interest tags and research relationship tags change based on research achievement information. According to statistics, among 20 Chinese scientific journals in 2010, the average publication cycle for 3,164 papers was 11.6 months. Therefore, researcher achievement information should be collected regularly. According to the process in Figure 1, extract keywords and co-authors from new achievement information as new interest tags and research relationship tags, calculate their weights using the method in Section 2.3.2, and compare them with existing tags. If tag content or weights have changed, replace the original data with new data and then visualize; if no changes occurred, directly visualize the original tags.

2.5 Researcher Profile Visualization

Researcher profiles can be viewed as tag clouds of user information. Based on tag weights, we can visually present researcher information using different sizes. Mature tools already exist for tag visualization, such as Wordle, tagCloud, Tagul, and Tagxedo.

3 Case Validation

To demonstrate the feasibility of the method, this paper constructs profiles for both domestic and international researchers. The domestic researcher is Professor YYT, an expert in microelectronics who has published hundreds of papers in domestic and international academic journals and important academic conferences. The international researcher is Professor P. Domingos from the University of Washington's Department of Computer Science and Engineering, a co-founder of the International Machine Learning Society whose academic level is recognized by peers worldwide. This paper provides the detailed construction process for Professor YYT; Professor Domingos's profile construction follows the same process, with both final profiles presented.

3.1 Data Collection

YYT's demographic attribute data comes from Baidu Baike and his personal homepage. As the world's largest Chinese encyclopedia, Baidu Baike contains over 15 billion entries covering nearly all knowledge domains. Researchers'

personal homepages also include introductions to their basic information and research content. We use web crawlers to obtain demographic attribute data.

Research achievement data sources include Chinese and foreign academic databases. Chinese academic databases mainly include CNKI, Wanfang, and VIP, which have overlapping content but complement each other. Foreign databases mainly include Web of Science and EI. The National Natural Science Foundation of China website contains information about researchers' recent projects, and the National Science and Technology Report Service System provides special reports, progress reports, final reports, and organizational management reports on research activities. A total of 892 journal paper records were collected (502 Chinese, 390 foreign), 120 conference paper records (6 Chinese, 114 foreign), 519 Chinese patent records, and 208 master's and doctoral thesis records.

Research social data comes from ResearchGate, one of the world's most successful academic social networking sites. We used the Octoparse web crawler tool to extract user friend relationships from this site. Although the site contains basic user information and achievement information, for data uniqueness we only obtained YYT's friend list from this site.

Similarly, we collected various types of information for Professor P. Domingos. Demographic attribute data came from his personal homepage, achievement data from Web of Science and EI databases (including 95 conference papers, 81 journal papers, and 4 monographs), and 59 research social data records from ResearchGate.

3.2 Data Preprocessing

In the collected achievement data, some data is not needed for profile construction. Therefore, we only extract required fields such as title, author, keywords, and publication date. In foreign data, publication dates use English formats; for data uniformity, we convert all date formats to "xxxx-xx-xx". Additionally, data duplication exists. Since our data is stored in databases, we use the DISTINCT command to remove duplicate records.

Key issues in data preprocessing include named entity recognition and name disambiguation.

3.2.1 Named Entity Recognition Some demographic attribute data is in text form, such as descriptions of researchers' work experience. We use the LTP named entity processing tool from Harbin Institute of Technology, calling its pyltp module through Python for named entity recognition, including names, places of origin, institutions, and birth dates. If we directly use Python's built-in segmentation dictionary, longer phrases like "西安电子科技大学" (Xidian University) would be segmented as "西安/电子/科技/大学", producing meaningless extractions. Therefore, we customize segmentation dictionaries to better extract researcher information.

3.2.2 Name Disambiguation Among YYT's English co-authors, name abbreviations like "Zhu ZY" and "Wang JY" appear. To better distinguish the real names corresponding to these abbreviations, we adopt Song Wenqiang's distribution clustering-based disambiguation algorithm for author name disambiguation in scientific literature. The algorithm's basic steps are: (1) Treat each document as a cluster, calculate similarity between any two documents to obtain an initial N×N matrix D; (2) Find the two most similar records in D and merge them into a new cluster; (3) Recalculate similarity between the new cluster and all other documents; (4) Repeat steps 2 and 3 until the final number of document clusters reaches the specified number. Match real researchers for each cluster based on Chinese co-authors. This method achieves 90% disambiguation accuracy. We implemented this algorithm in Python for name disambiguation. Table 4 statistics the name abbreviations and related literature quantities, which were distinguished using this algorithm. After disambiguation, among 94 "Zhu ZY" documents, 26 were identified as authored by "Zhu Zuoyun", 37 by "Zhu Zhenyu", and 13 by "Zhu Zhaoyi". Among 10 "Wang JY" documents, 3 were authored by Wang Jianyun and 2 by Wang Juyong. The distinguished author information was supplemented to corresponding attributes. Remaining undistinguished author information was further processed manually based on institutional information. Finally, processed data was stored in databases for subsequent tag extraction and weight calculation.

3.3 Model Representation and Tag Extraction

As described in Section 2.3.2, basic attribute dimension tags directly use information from the researcher information database. Research achievement Achv tags use literature titles. Due to the large number of achievements, we use literature titles from the past three years as Achv tags.

Research preference tags use keywords from achievement information. Since researcher literature comes from Chinese and foreign databases, we convert all English keywords to Chinese before statistical analysis, manually correct them, and then use frequently occurring keywords as researcher research topic tags. We calculate weights using formulas (1) and (2), setting the harmonic factor $\lambda = 0.5$, assuming preference weight and decay weight have equal proportion in comprehensive weight. For decay weight half-life, we calculate based on the definition of keyword half-life. Researcher keyword half-life refers to how long it takes for half of the keywords used in a given year to be created. The calculation formula is $hl = A + (50\% - B)/C$, where A is the number of years elapsed for the year whose cumulative percentage is closest to 50%, B is the cumulative percentage for that year, and C is the annual percentage for the first year when cumulative percentage exceeds 50%. Using 2017 as the starting year, the year with cumulative percentage closest to 50% is 2008, giving $A = 9$, $B = 46.76\%$, and $C = 5.16\%$, resulting in $hl = 9.627$ years. Calculation results are shown in Table 5 .

Research relationship tags use co-author names. We integrate all authors from

the collected researcher achievement information, obtaining 1,085 persons with research relationships. We calculate each author's contribution degree R_u and research relationship strength R_{ui} using formulas (3)-(7). The results are shown in Table 6 . The research relationship model for this researcher can be represented as:

$$\text{Relation} = \{\langle \text{YYT}, 261.5498, 2.98\rangle, \langle \text{ZZM}, 105.0262, 2.88\rangle, \langle \text{YJJ}, 101.7000, 1.96\rangle, ...\}$$

Based on the research relationship model, we construct a researcher relationship network with R_{ui} as edge weights and R_u as node weights. We import label and weight information into Pajek to generate YYT's research relationship network (see Figure 3 [Figure 3: see original paper]). Clicking any node allows viewing that researcher's profile.

3.4 Researcher Profile Visualization

We use the visualization tool Tagul to import obtained labels and set label sizes based on weights. In current mature visualization tools and existing tag cloud research, label settings mostly depend on subjective preferences without unified standards. To make researcher profile presentation more aesthetically pleasing, we set label sizes within 10. Basic characteristics are important for understanding researcher information and change minimally, so we set basic characteristic label size to 10. Research preference and relationship label sizes are calculated using:

$$\text{Size} = \frac{\text{Weight of a certain tag}}{\text{Sum of weights of all tags}} \times 10$$

keeping all label sizes within 10. Finally, using the researcher's photo as background, we create the researcher profile, as shown in Figure 4 [Figure 4: see original paper]. Using the same method, we generate Professor P. Domingos's researcher profile, as shown in Figure 5 [Figure 5: see original paper]. In practical applications, different colors can be used to distinguish various types of labels for clearer information presentation.

Discussion

This method comprehensively describes researcher information from basic attributes, research interests, and research relationships, and presents it intuitively using tag cloud principles, enabling science and technology managers to quickly grasp researcher information during decision-making and effectively improve decision efficiency. We discuss several issues in the profile construction process:

In named entity recognition for researcher personal profiles, we used a dictionary-based method. However, many studies exist on named entity recognition. Our case validation uses dictionary-based recognition, where dictionary completeness is an important factor affecting extraction effectiveness. To improve extraction accuracy, we added long phrases like "西安电子科技大学" to the dictionary. However, manually adding words for each researcher wastes considerable time. When processing batch information extraction, machine learning algorithms such as CRF can be combined for extraction, which has achieved good results in named entity recognition.

For the keyword Chinese-English alignment issue in our case validation, existing research has achieved automated translation using rule-based, example-based, statistical, and neural network-based methods. In recent years, neural machine translation based on deep learning has emerged. However, the accuracy of different machine translation methods needs improvement. In our case, we used manual correction for accuracy. With continuous technological improvement, automated translation can be adopted for large-scale Chinese-English keyword alignment when processing large datasets.

This paper proposes a researcher profile construction method based on multi-source scientific and technological management data. The method addresses the problems of single-dimensional description and non-intuitive presentation in current research. However, different researchers emphasize different information. Basic researchers focus on the level and quality of their achievements and their impact in the field, while applied researchers focus on the transformation value of their achievements. Exploring researcher profiles for different types of researchers is an area for future expansion.

References

[1] Liu Xue, Liu Ying, Kang Yunting, et al. Evaluation framework for researchers in national research institutions[J]. Bulletin of Chinese Academy of Sciences, 2012, 27(5): 527-537.

[2] Wang Feiyue. Major changes in knowledge production methods and science and technology decision support: facing big data and open information for scientific and technological situation analysis and decision services[J]. Bulletin of Chinese Academy of Sciences, 2012, 27(5): 527-537.

[3] Hua Bolin. Application of scientific and technological information big data in intelligence research services[J]. Library and Information Service, 2017, 61(16): 150-156.

[4] Song Peiyan, Chen Baixue, Xian Xin. Construction and empirical research of scientific expert information semantic model[J]. Information Studies: Theory & Application, 2017, 40(9): 119-124.

[5] EuroCRIS[EB/OL]. [2016-02-15]. http://www.eurocris.org/.

[6] Moreira C, Wichert A. Finding academic experts on a multi-relational network using Dempster-Shafer theory and Shannon's entropy[J]. Expert Systems with Applications, 2013, 40(14): 5740-5754.

[7] Yang L, Hu Z, Long J. Service of searching and ranking in a semantic-based expert information system[C]//Proceedings of the IEEE Asia-Pacific services computing conference. Los Angeles: IEEE Computer Society, 2010: 609-614.

[8] Duong TH, Uddin MN, Nguyen CD. Personalized semantic search using ODP: a study case in academic domain[C]//Proceedings of the international conference on computational science and its applications. Vietnam: Springer Berlin Heidelberg, 2013: 607-619.

[9] Li Gang, Ye Guanghui. Research on multi-source expert feature information fusion[J]. New Technology of Library and Information Service, 2014(4): 27-33.

[10] Wang Yuefen, Wang Xuefen, Yang Xiaoxiao. Construction scheme and process design of science and technology consulting expert database based on social network[J]. Journal of the China Society for Scientific and Technical Information, 2012, 31(2): 116-125.

[11] Lu Wei, Han Shuguang. Design and implementation of organization expert retrieval system[J]. Journal of the China Society for Scientific and Technical Information, 2008, 27(5): 657-663.

[12] Tang J, Yao L, Zhang D, et al. A combination approach to Web user profiling[J]. ACM transactions on knowledge discovery from data, 2010, 5(1): 1-44.

[13] Hua Bolin, Li Guangjian. Discussion on theory and application of multi-source information fusion in big data environment[J]. Library and Information Service, 2015, 59(16): 5-10.

[14] Cooper A. The Inmates Are Running the Asylum[M]. Translated by DING C. Beijing: Publishing House of Electronics Industry, 2006.

[15] Qiu X, Zhang Q, Huang X. FudanNLP: a toolkit for Chinese natural language processing[C]//Proceedings of the meeting of the Association for Computational Linguistics: system demonstrations. Sofia: the Association for Computational Linguistics, 2013: 49-54.

[16] Zhou L, Zhang D. NLPIR: a theoretical framework for applying natural language processing to information retrieval[J]. Journal of the American Society for Information Science & Technology, 2003, 54(2): 115-123.

[17] Liu Ting, Che Wanxiang, Li Zhenghua. Language technology platform[J]. Journal of Chinese Information Processing, 2011, 25(6): 53-62.

[18] Song Wenqiang. Name disambiguation and entity linking for authors in scientific literature[D]. Harbin: Harbin Institute of Technology, 2012.

[19] Zhang Shunrui, You Hongliang. Chinese name disambiguation based on hierarchical clustering algorithm[J]. New Technology of Library and Information

Service, 2010(11): 64-68.

[20] Zhu Yunxia. Research on author name disambiguation in Chinese bibliographic data[J]. Library and Information Service, 2014, 58(23): 143-148, 142.

[21] Cheng Y, Qiu G, Bu J, et al. Model bloggers' interests based on forgetting mechanism[C]//Proceedings of the international conference on World Wide Web. New York: ACM, 2008: 1129-1130.

[22] Fan Yujing. Reputation allocation for co-authors[J]. Journal of Intelligence, 1997(1): 37-38.

[23] Zhao Shuqing, Liu Yongsheng. Investigation on publication delay of 20 scientific journals in 2010[J]. Acta Editologica, 2011, 23(6): 491-493.

[24] WordArt.com WordCloudArtCreator[EB/OL]. [2018-01-30]. https://wordart.com/.

[25] Tagxedo WordCloud with Styles[EB/OL]. [2018-01-30]. http://www.tagxedo.com/.

[26] Chen Mupei. Measurement and aging research of academic keyword half-life[J]. Science and Technology Entrepreneurship Monthly, 2010, 23(8): 156-157.

[27] Xu Jianzhong, Zhu Jun, Zhao Rui, et al. Aerospace named entity recognition based on CRF algorithm[J]. Electronic Design Engineering, 2017, 25(20): 42-46.

[28] Xie Zhining. Research on Chinese named entity recognition algorithm[D]. Hangzhou: Zhejiang University, 2017.

[29] Zeng Guanming. Research on Chinese named entity recognition based on conditional random fields[D]. Beijing: Beijing University of Posts and Telecommunications, 2009.

[30] Liu Ying, Jiang Wei. Statistical machine translation based on translation rules[J]. Computer Science, 2013, 40(2): 214-217.

[31] Liu Zhanyi, Li Sheng, Liu Ting, et al. Improving example-based machine translation using statistical collocation models[J]. Journal of Software, 2012, 23(6): 1472-1485.

[32] Zhao Jing. Research on statistical Chinese-English machine translation technology[J]. Electronic Design Engineering, 2016, 24(21): 69-71, 75.

[33] Ding Liang, He Yanqing. Research on domain adaptation of machine translation integrating domain knowledge and deep learning[J]. Information Science, 2017, 35(10): 125-132.

[34] Li Jingxuan. Research on statistical machine translation model based on deep neural networks[D]. Harbin: Harbin Institute of Technology, 2016.

[35] Yang Nan. Research on statistical machine translation based on neural network learning[D]. Hefei: University of Science and Technology of China, 2014.

Author Contributions

Fan Xiaoyu: Proposed initial research ideas and drafted the paper;
Dou Yongxiang: Guided overall research ideas and revised the paper multiple times;
Zhao Pengwei: Guided research ideas;
Zhou Xiao: Revised the paper.

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