Research on the Construction of a Label System for Scientific Research Institutions

Xin Xian[1] and Chuan Ge[2] and Huiying Yang[1] and Bo Yang[3]

[1] The Institute of Scientific and Technical Information of China, Beijing 100038, China
[2] Shanxi Provincial Center for Science and Technology Intelligence and Strategic Research, Shanxi 030024, China
[3] Center of science and technology review and assess, Beijing 101117, China
xianxin@istic.ac.cn

Abstract. Research institutions, as the core entities in research activities, constitute a vital component of technological prowess. The intuitive and dynamic revelation of information pertaining to research institutions holds significant implications. Addressing issues such as the uni-dimensional depiction of research institutions, coarse granularity of profile labels, and the lack of a clear hierarchical structure, this study employs research institution attributes and relevant data to construct a comprehensive profile label system across four dimensions: basic attributes, research attributes, domain attributes, and relational attributes. Leveraging LDA topic modeling and social network analysis methods, the study generates model labels for research institutions. Using a specific research institute as an example and drawing on diverse and heterogeneous data from sources such as the official institution website, literature databases, patent databases, project databases, and information media platforms, the feasibility of this approach is validated. The results underscore the stability and scalability of the research institution profile label system.

Keywords: Research Institutions, Label System, Institutional Profile.

1 Introduction

Research institutions refer to organizations engaged in scientific and technological research and development, primarily including research institutes, research-oriented universities, and R&D enterprises. They constitute the core entities in research activities and play a crucial role as essential components of technological prowess, exerting significant influence in propelling national scientific and technological development. A comprehensive, dynamic, and intuitive revelation of multidimensional information about research institutions contributes to self-awareness, benchmarking, collaboration, and enables third parties to understand and predict institutional behavior. This, in turn, assists management in making macro-level decisions.With the development of internet technology, the permeation of open access principles, and the widespread adoption of social media, traces of activities originating from research institutions, such as research

C. Bai et al. (eds.), *Proceedings of 2023 China Science and Technology Information Resource Management and Service Annual Conference (COINFO 2023)*, Advances in Economics, Business and Management Research 293,
https://doi.org/10.2991/978-94-6463-498-3_11

outcomes, academic conferences, development trends, and research data, are dispersed across various channels, including internal institutional sources, public websites, commercial databases, and social media. These traces manifest in diverse modalities, such as structured data, plain text, images, and audiovisual content, presenting characteristics of large volume, rapid updates, and diverse structures. While accumulating a wealth of materials for research and practical applications related to research institutions, this diversity also poses challenges in gaining an intuitive, comprehensive, fast, and accurate understanding of these institutions.

Constructing a profile of research institutions based on an organizational tagging system can address this issue to some extent. Tags are succinct textual representations, manually defined or generalized, of the characteristic properties of a target object. They serve as symbolic expressions of the target object's features, providing an organizational perspective on data from a business application viewpoint. Tags possess characteristics of semantic clarity and lack of ambiguity, facilitating understanding and subsequent retrieval, processing, and application of information. As the number of tags grows, it becomes necessary to establish a tagging system to classify and manage them for ease of querying and application [1]. However, existing organizational tagging systems face challenges such as coarse granularity and a lack of clear hierarchical structures. In this study, we integrate relevant theories and technologies related to profile tagging, focusing on research institutions as the target objects. We aim to construct a multi-dimensional and multi-level tagging system for research institution profiles. This approach transforms complex and heterogeneous organizational information into semantically rich and structured tags, laying the foundation for the development of an intuitive and rapid profile of research institutions.

2 Related work

In the context of the internet environment, the concept of user profiles originated from Alan Cooper's introduction of "personas" [2], which are hypothetical prototypes representing real individuals [3]. With the advent of the big data era, the term "User Profile" gradually became mainstream, indicating a collection of labels describing users based on accumulated user data [4]. Unlike traditional data resource descriptions and organizational methods, profiles organize data resources in a tag format, making them readable, easily understandable, and user-friendly. The earliest appearance of profiles was in user-centric profile systems, designed to describe and reveal various dimensions of user characteristics. These systems were employed for tasks such as precision marketing, churn prediction, and new customer identification. Over time, the scope of profiling expanded to encompass groups, enterprises, industries, and even cities – essentially any object that requires description.

2.1 Tag Construction Methods

Early popular methods for tag construction were predominantly based on the collaborative tagging concept proposed by Golder [5]. Tomas Vander Wal combined "folk" and "taxonomy" to introduce "Folksonomy," characterized as a "bottom-up social classification." In this approach, network users spontaneously add tags to certain information, ultimately selecting high-frequency tags as the classification names for that information [6]. Subsequently, Cai et al. proposed using both structured and unstructured tags with different weights from collaborative tagging systems to construct a tag system [7]. Collaborative tagging represents not only the unique insights of network users but also the common understanding shared by all network users. While collaborative tagging is highly flexible and cost-effective, as it does not require pre-learning or training, the lack of control over tag wording leads to issues of unclear meaning and unclear hierarchical structure in the final tag system.

With the evolution of ontology, research on constructing tag systems using ontology has gradually increased. Maleszka et al. utilized existing ontology frameworks to manually or automatically construct ontologies to generate tag systems [8]. However, tag systems derived from ontology currently suffer from low accuracy, redundancy, or missing tags, structural issues, and significant human involvement in ontology construction, leading to high costs. Some scholars have employed model algorithms to construct tag systems. For instance, Farseev et al. [9] used a topic modeling approach to build tag systems, but this type of tag system tends to have coarse granularity, hindering precise profiling.

2.2 Tag Construction Methods for Institutional Profiling

In the realm of institutional profiling, current efforts have primarily focused on enterprise profiles utilized in e-commerce, risk assessment, and market supervision, with relatively limited attention to research institution profiling. ChunR et al. introduced the concept of enterprise image, discussing five main dimensions (comfort, enterprising, strict, relaxed, competence) and two subordinate dimensions (hobbies, arbitrary judgments) used to evaluate the impact of enterprise reputation on employees and customers [10]. Current enterprise profile tag systems primarily encompass dimensions such as basic attributes (enterprise size), credit attributes, operational attributes (profitability, product promotion), financial attributes, research investment (product development), corporate culture, industry attributes, financing attributes, internal and external relations features, and evaluation information (product reputation) [11-13] These dimensions support enterprise security risk assessment and regulatory compliance.

In the context of research institutions, scholars have also embarked on explorations. Johannesson et al. proposed an institutional conceptual description model that describes institutions from the perspectives of roles, rules, rights, responsibilities, and processes, defining criteria for describing different entities [14]. Meng Lin [15] focused on dynamic attribute profiling within institutional profiles, conducting feasibility studies and implementations in community discovery, relationship extraction, and institutional interest discovery. Guo Hongmei et al. [16], based on an ontology model, defined and described the characteristics of research institutions in terms of basic attributes, social attributes, research attributes, and relationship attributes. They constructed an institutional tag system with dimensions of descriptive information tags, related relationship tags, and related institution tags across three dimensions.

Scholars both domestically and internationally have attempted to construct institutional profile tags and tag systems by leveraging description models, ontologies, or mining algorithms, accumulating valuable theoretical and practical experiences. However, several issues persist: (1) Application-Centric Focus: The majority of existing research is primarily oriented towards specific applications, with a singular descriptive perspective. There has been a lack of comprehensive analysis of the overall attributes and relationships of research institutions. (2) Coarse-Grained and Unclear Hierarchical Structures: Research institutions exhibit rich and complex attribute features and relational structures. Existing profile tag systems mostly possess coarse granularity, with unclear hierarchical structures. This results in poor scalability, significantly impacting subsequent application services tailored for different scenarios. To address these challenges, this study comprehensively analyzes the attribute features and relationships of research institutions based on their behavior and relevant data. Utilizing an enumerative classification method, the study aims to construct a set of tag systems and a profiling construction process that can accurately define and describe the attribute features of research institutions.

3 Research Methodology

The construction of a research institution portrait involves the three-dimensionalization and conceptualization of flattened institutional data [17]. This process entails the application of various data analysis methods to process institution-related data, extracting highly generalized labels to form a research institution portrait. The research institution portrait model consists of three layers: the data layer, analysis layer, and application layer, as illustrated in Figure 1. The data layer encompasses various institutional data resources, including traditional literature data such as papers, patents, books, and reports published by research institutions. Additionally, it includes open-source data from institution websites, core media, mainstream social media, and other sources. The analysis layer involves the application of various analytical methods to transform raw data into portrait labels. The application layer focuses on conceptualizing institution labels for specific application scenarios to generate a comprehensive research institution portrait.

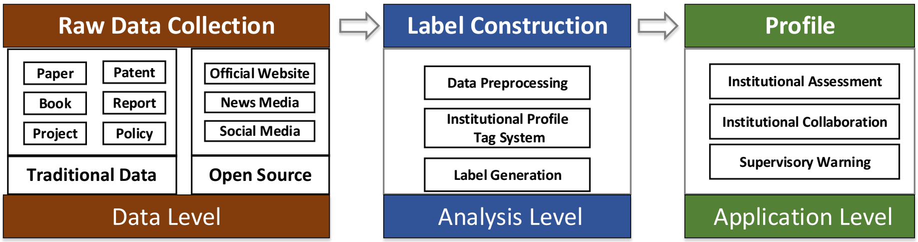

Fig. 1. Process of Constructing a Research Institution Profile.

3.1 Attribute Feature Analysis

To construct a research institution portrait, a profound analysis of the research institution entities is essential for obtaining relevant information. Through a multidimensional analysis, synthesis, and abstraction of the institution itself, its behaviors, internal and external aspects, as well as static and dynamic attributes, it becomes apparent that a research institution is a social entity with rich attribute features and interrelationships [16]. In addition to possessing the fundamental characteristics, economic features, legal features, and other attributes common to organizations, research institutions exhibit unique research attributes. These include various multimodal entities such as institutions, personnel, teams, projects, patents, and more. Moreover, they entail heterogeneous relationships such as competition, collaboration between institutions, publication of research outcomes, and dissemination of information, as illustrated in Figure 2.

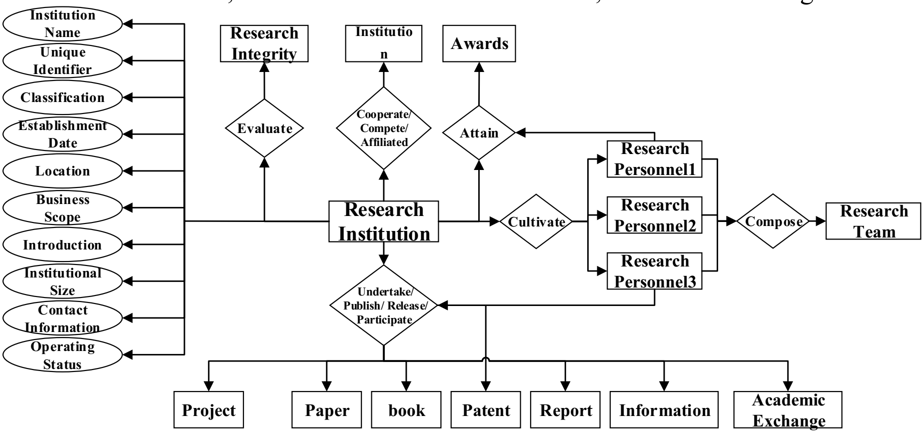

Fig. 2. Conceptual Model of Scientific Research Institutions.

Basic information about research institutions is relatively stable and easily accessible, encompassing various forms of institution names, institution types, establishment times, and geographical locations. Research features manifest in scientific activities, involving undertaking/participating in research projects to secure research funding. Research outcomes are disseminated in the form of academic papers, scholarly monographs, patents, and technological reports. Research institutions also engage in subsequent experimentation, development, and application of scientific achievements. Typically bearing the responsibility of nurturing talent for the nation, research institutions commonly operate in team collaboration, ranging from long-term laboratory setups for major tasks to small groups formed for temporary assignments. In the process of applying for research projects, collaborative relationships are formed between units undertaking and participating in the same project, while significant competition emerges among applying units for the same guideline. Research institutions establish subordinate organizations/departments based on development goals and role positioning. Therefore, constructing a research institution portrait relies on data from multiple sources, including official institution websites, domestic and international literature databases, social media, news websites, and others, as detailed in Table 1.

Table 1. Examples of Data Sources for Scientific Research Institutions.

Data Type	Main information content	Data source
Basic Information	Institution Name, Institution Type, Establishment Time, Geographical Location, Institution Introduction, Contact Information	Institution Official Website, Institution Business Database
Research Activities	Research Projects	Domestic and International Project Databases, Science and Technology Management Information System, Institutional official websites, etc.
	Academic Papers	Domestic and International Paper Databases, Institutional official websites, etc.
	Science and Technology Reports	Domestic and International Science and Technology Report Systems
	Patents	Domestic and International Patent Retrieval Databases
	Achievement Transformation	Technology Achievement Transformation System
	Academic Exchange	Institutional official websites, News Website, Social Media, etc.
	Information	Institutional official websites, News Website, Social Media, etc.

3.2 Construction of Label System

The labeling system serves as a crucial foundation for the portrait, with its breadth and depth directly influencing the portrait's outcome. In this study, a comprehensive approach combining the rank enumeration classification method and the facet-assembly classification method is employed to construct the labeling system for research institutions. The rank enumeration classification method is based on knowledge classification, utilizing conceptual division and generalization principles to create a hierarchical tree structure that categorizes and details various concepts and aspects of entities.

Based on the content of research institution attributes, the rank enumeration classification method is applied to categorize the first and second-level labels of the research institution portrait. This results in the formation of a four-dimensional tree structure framework, including basic attributes, domain attributes, research attributes, and relationship attributes, as illustrated in Figure 3.

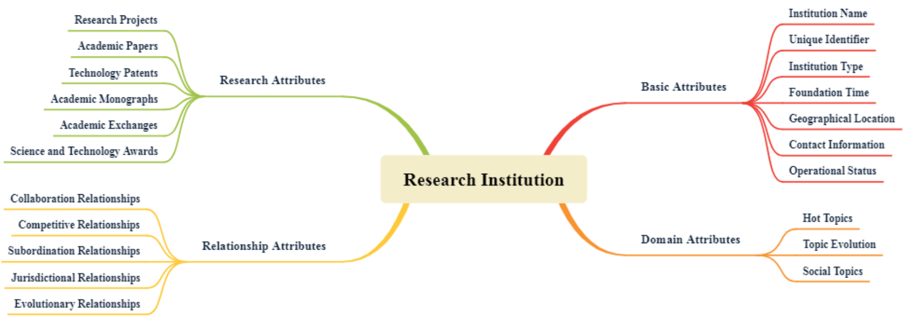

Fig. 3. Conceptual Model of Scientific Research Institutions.

Basic attributes encompass relatively stable static information about research institutions. On the basis of the first and second-level categories, the facet-assembly classification method is applied to further divide them into the smallest granularity, all of which are factual labels. For research attributes, domain attributes, and relationship attributes, a detailed analysis is performed on the second-level categories based on factual labels, statistical labels, and model labels. Factual labels involve no complex computational logic; they only require extracting raw data to generate labels. Statistical labels necessitate statistical calculations and divisions of specific data to produce different label values. Model labels generally require complex calculations through algorithms/models [18]. Research attributes depict the primary research activities of research institutions, reflecting their research capabilities to some extent. These labels cover three types, as shown in Table 2. Domain attributes reflect the main research areas of research institutions, revealed by extracting research topics from various research activities. Therefore, it requires model-based calculations, as shown in Table 3. Relationship attributes describe the network of relationships between institutions, mainly revealed through factual labels and model labels, as shown in Table 4.

Table 2. Scientific Research Institution's Attributes Portrait Tag System.

First-level Category	Second-level Category	Labels		
		Factual Labels	Statistical Labels	Model Labels
Research Attributes	Research Project	Project Name	Total Number of Projects, Project Annual Distribution, Project Level Distribution	Project Theme, Theme Intensity, Theme Network
	Academic Paper	Paper Title	Total Number of Papers, Number of Journal Papers, Number of Conference Papers, Number of Chinese Papers, Number of Foreign Language Papers, Number of	Paper Topic, Theme Intensity, Theme Network

			Highly Cited Papers, Number of High-Quality Papers, Paper Annual Distribution	
	Technology Patents	Patent Title	Total Number of Patents, Patent Annual Distribution, Patent Domain Distribution	
	Technology Report	Report Title	Total Number of Reports, Report Annual Distribution	
	Academic Monograph	Monograph Title	Total Number of Monographs, Monograph Annual Distribution	
	Academic Exchange	Meeting Title	Number of Exchanges, Annual Distribution of Exchanges	Conference Theme Distribution

Table 3. Field Attribute Portrait Tag System of Scientific Research Institutions.

First-level Category	Second-level Category	Model Labels
Domain Attributes	Hot Topics	Research Topics, Theme Intensity, Theme Network
	Topic Evolution	Research Topics, Theme Intensity, Research Time, Theme Temporal Network
	Social Topics	Topic, Topic Intensity

Table 4. Relationship Attribute Portrait Tag System for Scientific Research Institutions.

First-level Category	Second-level Category	Labels	
		Factual Labels	Model Labels
Relationship Attributes	Collaborative Relationships	Collaborating Institutions, Type of Collaboration	Collaboration Intensity, Collaboration Themes, Collaboration Network
	Competitive Relationships	Competing Institutions, Type of Competition	Competitive Themes
	Affiliation Relationship	Superior Organization	

	Jurisdictional Relationship	Subsidiary Organization
	Evolutionary Relationship	Modification Time, Modification Type, Organization Before Change, Organization After Change

3.3 Label Generation

Label generation, alternatively known as tag extraction or tagging, is the cornerstone of profile construction. The basic information of institutions in the original dataset is predominantly structured data, characterized by its conciseness and lack of ambiguity, and can thus be directly repurposed as fundamental attribute labels for these institutions. Labels pertaining to research attributes, which are statistical in nature, can be derived through mathematical formulas; this process will not be further detailed here. This section primarily discusses the methodologies for generating labels for research themes and relational networks.

Multi-source Data Fusion. Institutional profiling that integrates multi-source heterogeneous data can more comprehensively reveal the full spectrum of an organization. Additionally, cross-verification of information from multiple sources can reduce information omissions and enhance the credibility of the profiling results. Institutional profiles in scientific research involve multi-source heterogeneous data such as research projects, academic papers, patents, scientific reports, and news information. These data sources may play different roles in constructing profiles for various attributes of an institution and should be assigned different weights, namely:

$$\text{Institution}=\alpha A+\beta B+\gamma C+\cdots$$

In the formula, A, B, C, etc., represent profile data from different sources, such as projects, papers, patents, etc.; α, β, γ, etc., denote the respective weights assigned to these multi-source data.

Research Topic Identification. The Latent Dirichlet Allocation (LDA) topic model is a probabilistic topic model for discrete data sets (such as document collections), proposed by David Blei, Andrew Ng, and Michael I. Jordan in 2003. It identifies latent topic information in documents, assuming each document is composed of a mix of multiple topics, and each topic in turn is composed of a distribution of various words. This study employs the LDA topic model for profiling analysis based on datasets comprising titles, keywords, abstracts, and other textual information from projects, papers, patents, and news. The LDA topic model algorithm is implemented using the gensim toolkit in Python.

Collaborative Network Mining. Scientific collaboration contributes to the complementarity of strengths among research institutions, promotes resource sharing, and enhances the quality of scientific research. Institutional scientific collaborations are manifested in phenomena such as co-authorship of papers and project cooperation. This study applies social network analysis methods to depict and analyze the collaborative relationship networks of institutions. In this network, institutions are treated as nodes, and the cooperative relationships between them form the connections between these nodes.

3.4 Profile Visualization

Profile visualization encompasses various methods such as statistical graphs (histograms, line charts, etc.), tag clouds (word clouds), and network maps. Statistical graphs are suitable for statistical labels; tag clouds display labels in varying font sizes or colors based on their weights, suitable for showcasing research hotspots and the like, with mature tools available like Word Art, WordSift, TagCrowd, ToCloud, or using the Python WordCloud package for coding, such as Wordle, tagCloud, Tagul, Tagxedo, etc. Network maps are appropriate for displaying various cooperative relationships, with network visualization facilitated by tools such as citespace, Gephi, VOSviewer, and others.

4 Empirical Research

To validate the effectiveness of the research institution profile tag system and construction methodology, Institution Z was selected for a case study. This research institution has long been engaged in scientific and technological information analysis, data governance, and providing decision-making support to government departments. The study constructs an institutional profiling application under the scenario of institutional assessment, based on four types of data from Institution Z: basic information, paper data, project data, and news data. This involves two main aspects: research attributes and field attributes (as shown in Table 5). The former is primarily used for quantifying research contributions, while the latter mainly verifies the alignment with the institution's role and responsibilities.

Table 5. Relationship Attribute Portrait Tag System for Scientific Research Institutions.

First-level Category	Second-level Category	Labels		
		Factual Labels	Statistical Labels	Model Labels
Research	Research Project	Project Name	Total Number of Projects, Project Annual Distribution	Project Theme, Theme Intensity

Attributes	Academic Paper	Paper Title	Number of Journal Papers, Paper Annual Distribution	Paper Topic, Theme Intensity
	Technology Patents	Patent Title	Total Number of Patents, Patent Annual Distribution	
Domain Attributes	Hot Topics			Research Topics, Theme Intensity
	Topic Evolution			Research Topics, Theme Intensity
	Social Topics			Topic, Topic Intensity

4.1 Data Collection

The basic attribute data of Institution Z was sourced from an institutional database, including structured data such as the institution's name, unique identifier, type, foundation date, geographic location, summary, and contact information. For paper data, the Wanfang database was chosen, with a search conducted on November 1, 2023. The search expression was "(Author Affiliation:(Institution Z)) and Publication Date: 2013-01-01 TO 2023-11-01," focusing on "Journal" document types, yielding a total of 3862 entries. The project data was sourced from the National Social Science Fund Project Database, capturing all social science fund projects of Institution Z over the past ten years, amounting to 38 projects. For patent data, the Zhihuiya Patent Database was used to gather all patents of Institution Z, totaling 100 items. The news data was collected using a commercial data crawling platform, capturing news media articles featuring "Institution Z" published from May 1, 2023, to October 31, 2023, totaling 3827 articles.

4.2 Data Preprocessing

The raw data collected from various channels may have issues like missing data, duplicates, inconsistencies, and homonyms with different meanings, which can affect the quality of the profile. It is necessary to clean and process the data according to specific needs to maximize its value.

The quality of social science project data and patent data is relatively high, allowing direct use for subsequent tag extraction. Of the 4374 papers collected, 1247 were not related to Institution Z, leaving 3127 valid papers. The institutional names among these papers varied in format, necessitating standardization: firstly, replacing names of secondary and tertiary institutions with the full name of the primary institution, and secondly, unifying different expressions of the same institution with its full name, including abbreviations and aliases, identified from the author affiliation field. The news data contained a substantial amount of duplicate content, including direct reprints and similar information. After removing duplicates, links with expired original articles, and selecting data published by core media, 1571 articles were considered valid for the experiment.

4.3 Application of Research Institution Profiling in Institutional Assessment Scenarios

The construction of research institution profiles holds significant practical value for the public, research institutions, and government departments. Based on a research institution's profile tag system, appropriate attribute tags can be selected according to the application scenario, deriving various profiling services to achieve in-depth knowledge services. For example, targeting research institutions, analyzing profiles based on research attributes and domain attributes can precisely pinpoint institutional characteristics, identify strengths and weaknesses, and aid in institutional development. For government departments, utilizing multidimensional institutional profiles helps science and technology managers comprehensively and intuitively understand information about research institutions, objectively evaluate their levels, and provide reference for assessments, supervision, and scientific decision-making related to research institutions.

Based on the experimental data of Institution Z, the statistical tag profile of its research attributes is illustrated in the following figure. It shows that over the past decade, the overall scientific research output of the institution has been in decline, particularly evident in the area of journal publications. The institution undertakes a relatively small number of projects annually, with the peak of patent applications occurring in 2019 and significant fluctuations in the number of applications each year.

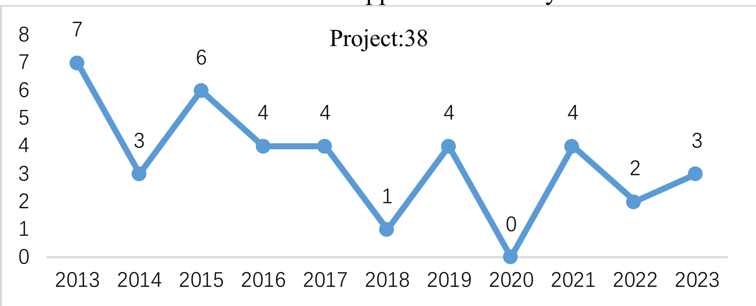

Fig. 4. Annual Distribution of Projects at Institution Z.

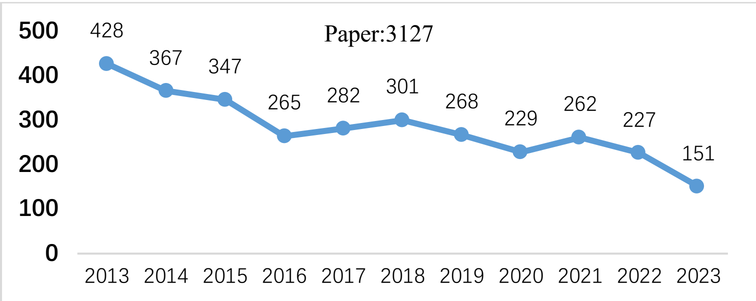

Fig. 5. Annual Distribution of Publications by Institution Z.

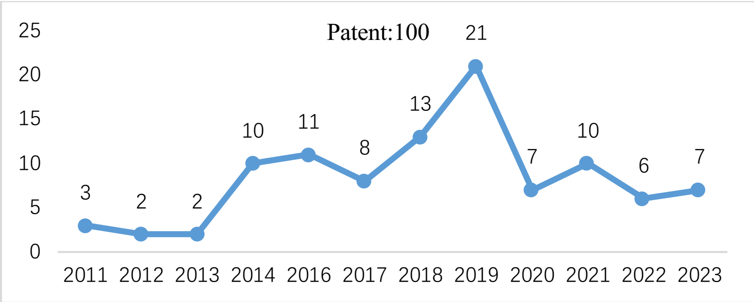

Fig. 6. Annual Distribution of Publications by Institution Z.

To verify the alignment of Institution Z's development with its designated responsibilities, academic papers from the past two years and news data from the last six months were selected as experimental data. The Latent Dirichlet Allocation (LDA) topic model was used to generate relevant topic tags, which were then compared against the institution's responsibility profile. The research themes of the academic papers primarily focus on areas such as artificial intelligence, competitive intelligence, open science, technological innovation, scientific data, international cooperation, science policy, and scientific intelligence (as shown in Figure 7). These themes fall within the scope of the institution's designated responsibilities.

Fig. 7. Annual Distribution of Patents by Institution Z.

As shown in Figure 8, the overall development trend of Institution Z's social impact from May 1, 2023, to October 31, 2023, peaked in July 2023. Textual analysis reveals that in July, Institution Z released a representative industry report and organized and participated in significant academic conferences.

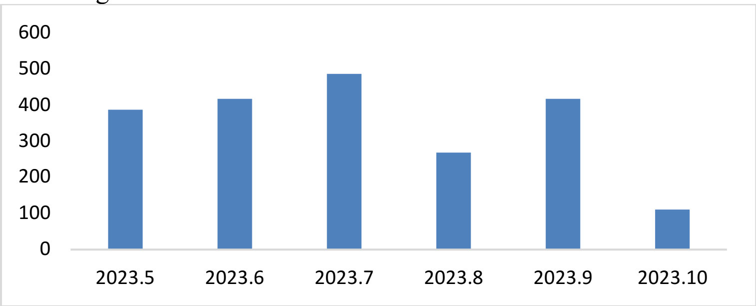

Fig. 8. Annual Distribution of Patents by Institution Z.

Through the thematic analysis of 1571 news articles (as illustrated in Figure 9), it is evident that social media and netizens primarily focus on Institution Z's activities in areas such as artificial intelligence, scientific papers, models, and information services. This aligns well with the content pertinent to Institution Z's designated responsibilities.

Fig. 9. Hot Topics in News and Information for Z Organization

5 Discussion

This study constructs a profile tag system for research institutions based on their attribute characteristics and relevant data, encompassing four dimensions: basic attributes, research attributes, field attributes, and relationship attributes. The system includes three types of tags: factual, statistical, and model tags. Methods such as statistical graphs and tag clouds are utilized to display information about research institutions. Discussion on Some Issues in the Construction Process of the Research Institution Profile Tag System:

Complex Nature of Research Institutions: The analysis of research institutions' attribute characteristics reveals that they are complex entities involving various types like institutions, personnel, teams, projects, and papers. Personnel and teams, being active entities in research with rich attributes, are considered to have their separate profiling categories. Therefore, they are not included in the research institution profile tag system. This reflects an understanding of the distinct nature of human elements in research settings.

Focus on Institutional Profile Tag System: The primary focus of this study is on the construction and application of the institutional profile tag system. It does not delve extensively into the cross-verification and fusion of multi-source data. Consequently, in the empirical research, only one data source was chosen for each type of data, such as selecting only the Wanfang database for paper data. This approach simplifies the data handling process while still providing valuable insights.

Topic Analysis in Profile Tag System: The study involves different dimensions of topic analysis within the research institution profile tag system. Existing research has addressed the issue of topic identification, with common methods including probabilistic graphical models, matrix decomposition, word embedding, deep learning, clustering, and frequency statistics. These methods show varying performance across different scenarios and datasets. To ensure better interpretability of the topic identification results, the study opted for a probabilistic graphical model approach. Manual correction methods were used in the empirical study, and with the increase in data volume, more topic identification methods can be explored in the future. This adaptability ensures the relevance and accuracy of the profiling process.

6 Conclusion

In response to the issues identified in existing research, such as a unidimensional perspective in describing research institutions, lack of intuitive display, coarse granularity of portrait labels, and unclear hierarchical structures, this study constructs a portrait label system for research institutions based on their attribute characteristics and relevant data. This system encompasses four dimensions: basic attributes, research attributes, domain attributes, and relational attributes. The study introduces a model label generation method for research institutions based on topic recognition models and social network analysis techniques. The feasibility of this method was validated using a domestic research institution as a case study. This approach effectively addresses the problems encountered in the description and display processes of research institutions, enhances the data services of institutions in various scenarios, and mitigates the application gap caused by the heterogeneity of multi-source data.

However, this study has certain limitations: firstly, research institutions can be categorized into research institutes, research-oriented universities, and R&D-oriented enterprises based on their nature, and into basic research institutions, applied research institutions, and public welfare research institutions based on the nature of their research. It is necessary to further investigate the characteristic differences among these types of research institutions and refine the label system for research institution portraits; secondly, the construction of portrait labels for research institutions is an open and iterative process, which requires continuous improvement of the label system and updating of label data in response to changes in original data, label generation methods, and business scenarios.

Acknowledgements This research received the financial support from Shanxi Province Science and Technology Cooperation and Exchange Special Project under grant numbers of 202204041101034 and State Grid Corporation of China Headquarters Science and Technology Project of 1400-202357341A-1-1-ZN, respectively. Our gratitude also goes to the anonymous reviewers for their valuable suggestions and comments.

References

1. Yinzi, R., Lele, J.: Label Hierarchies: Business Oriented Data Asset Planning Methodology. China Machine Press, Beijijng China (2021).
2. Cooper, A.: The inmates are running the asylum. Indianapolis, IN Sams (1999).
3. Cooper, A., Chris, D.: The inmates are running the asylum : why high-tech products drive us crazy and how to restore the sanity. Publishing House of Electronics Industry, Beijijng China (2006).
4. Yahui, Z., Fanglin, L., Lin, L.: A Review of User Profile Research in the Context of Big Data: Knowledge System and Research Prospects. Research on Library Science 1(24), 13-24 (2019).
5. Golder, S., Huberman, B.: Usage Patterns of Collaborative Tagging Systems. Journal of Information Science 1(32), 198-208 (2006).
6. Cuiying, W.: The Comparative Study of the Ontology and Folksonomy. Library Development 1(5), 85-88 (2008).
7. Cai, Y., Li, Q., Xie, H.: Exploring personalized searches using tag-based user profiles and resource profiles in folksonomy. Neural Networks: The Official Journal of the International Neural Network Society 1(58), 98-110 (2014).
8. FerreiraSatler, M., Romero, F. P., Olivas, J. A., SerranoGuerrero, J.: Fuzzy ontology-based approach for automatic construction of user profiles. In Rough Sets and Current Trends in

Computing: 9th International Conference, RSCTC 2014, Granada and Madrid, Spain, July 9-13, 2014. Proceedings 9 pp. 339-346. Springer International Publishing (2014).
9. Farseev, A., Nie, L., Akbari, M.: Harvesting Multiple Sources for User Profile Learning: a Big Data Study. Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, 235-242 (2015).
10. Chun, R., Davies, G.: The influence of corporate character on customers and employees: Exploring similarities and differences. Journal of the Academy of Marketing Science 34(2), 138-146 (2006).
11. Juan, T., Dingju, Z., Wenhan, Y.: Research on Enterprise Portraits Based on Big Data Platforms. Computer Science 45(z2), 58-62 (2018).
12. Lou, W.: Research on Enterprise Portraits for Credit Reporting. Northwest University, (2019).
13. Xiaobin, H., Mingxin, Zhang.: Construction of Enterprise Competitor Portrait Based on Multi-source Data. Journal of Modern Information 40(11), 13-21,33 (2020).
14. Johannesson, P., Bergholtz, M., Eriksson, O.: Institution Aware Conceptual Modelling. In The ER Forum 2017 and the ER 2017 Demo Track, Valencia, Spain, November 6-9, 2017 pp. 263-270. CEUR-WS. Org (2017).
15. Lin, M.: Research on the Method of Institutional Portrait Based on Multi source Information Fusion. Beijing University of Posts and Telecommunications 1(11), (2018).
16. Hongmei, Guo., Jianxun, Zeng.: Profiling Attribution on the Label of Scientific Research Institutions Based on the Ontology Model. Journal of the China Society for Scientific and Technical Information 41(6), 574-583 (2022).
17. Xiwei, Wang., Ruonan, Jia., Yujiao, Sun.: Research on the Construction of Data-driven Social Network Public Opinion Polarized Group Portrait. Information and Documentation Services 42(6), 21-30 (2021).
18. Xiaomeng, Du., Xiaomi, Tang., Yinhu, Zhang.: Practical Guide for Big Data User Behavior Profile Analysis. Publishing House of Electronics Industry, Beijijng China (2021).
19. Blei, D. M., Ng, A. Y., Jordan, M. I.: Latent dirichlet allocation. Journal of machine Learning research 3(Jan), 993-1022 (2003).
20. Xiaoyu, Fan., Yongxiang, Dou., Bingwei, Zhao.: Research on the Method of Constructing Scientific Researcher Portraits by Integrating Multi source Data. Library And Information Service 62(15), 31-40 (2018).