ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-202304.00330

Research on Normalization of Institution Names in Chinese Bibliographic Data: Postprint

Authors: Yang Zhao, Ren Juan

Date: 2023-04-01T00:00:00+00:00

Abstract

[Purpose/Significance] In the big data era, institutional name data exhibits new characteristics such as massiveness, dynamism, and diversity. Institutional name normalization can improve data reliability in scientific research management, discipline evaluation, and discipline services within big data environments, and enhance the quality and application effectiveness of data retrieval based on institutional names. [Method/Process] From both linguistic and model construction perspectives, this study investigates institutional name normalization, constructs a framework model for institutional name normalization based on co-occurrence relationships and similarity, proposes a method for entity boundary recognition in institutional names, compiles a multi-level institutional vocabulary, presents an institutional name normalization method, and finally selects Chinese literature bibliographic data from 2008-2018 for experimental validation. [Results/Conclusion] The experimental results verify the effectiveness of the model and provide certain insights for the normalization of other types of institutional names.

Full Text

Research on Institution Name Normalization Based on Chinese Bibliographic Data

Yang Zhao[1], **Ren Juan**[2],[3] [1]Shanghai Jiao Tong University Library, Shanghai 200240 [2]Shanghai Publishing and Printing College, Shanghai 200093 [3]Shanghai Research Institute of Publishing and Media, Shanghai 200093

Abstract: [Purpose/Significance] In the era of big data, institution name data exhibits new characteristics such as massiveness, dynamism, and diversity. Institution name normalization can improve data reliability in scientific research management, discipline evaluation, and subject services under big data environments, and enhance the quality and application effectiveness of data retrieval

based on institution names. [Method/Process] This study investigates institution name normalization from both linguistic and model construction perspectives, proposes a framework model for institution name normalization based on co-occurrence relationships and similarity, presents an entity boundary identification method for institution names, compiles a multi-level institution vocabulary, and proposes an institution name normalization method. Finally, Chinese bibliographic data from 2008-2018 were selected for experimentation. [Result/Conclusion] Experimental results verify the effectiveness of the model and provide insights for normalizing other types of institution names.

Keywords: institution name; normalization; model construction; big data; entity boundary recognition

---

Institution names comprehensively reflect an institution's basic attributes, internal patterns, and particularities. Institution names include standard names, former names, translated names, merged names, and affiliated independent names, which can be categorized into standard names and variant names. Standard names refer to those officially released by authoritative bodies according to national standards, while variant names refer to multiple expressions of the same institution entity, mainly including full names and abbreviations, simplified and traditional Chinese forms, multilingual translations, cataloging errors, and names from different data sources and time periods. Institution name normalization aims to consolidate different expressions of the same institution entity, establish correspondence between standard names and variant names, and achieve institution identification by assigning unique identifiers. In essence, regarding the institution expansion retrieval functions of major literature databases, most provide "author + affiliation" filtering without distinguishing hierarchical relationships. Both WOS and SCOPUS databases offer institution expansion retrieval functions that enable first-level institution name normalization, while VIP, CNKI, and Wanfang databases provide author affiliation fields for journal articles. Among them, VIP annotates the correspondence between authors and institutions, CNKI does not, and Wanfang uses XML documents to annotate the correspondence between authors and first-level institutions. None of these databases address normalization below the second level.

Regarding research methods for institution name normalization, scholars have proposed rule-based and statistics-based approaches. The former primarily uses institution name ending identifiers to trigger boundary recognition, summarizing combination rules, semantic patterns, and grammatical features for identification through keywords. However, lacking semantic relationship definitions between various institution name types and relying solely on formal name matching leads to problems such as omissions and erroneous statistics. The latter mainly adopts the paradigm of combining text features with machine learning algorithms based on corpora. Nevertheless, machine learning algorithms are black boxes that lack interpretability for results, making it difficult to identify error causes and propose correction strategies when mistakes occur.

In terms of cataloging differences between Chinese and English institution names, Chinese institution names generally do not contain spaces or commas separating multi-level names as in English, requiring automatic word segmentation and named entity boundary recognition. This paper studies institution name normalization from linguistic and model construction perspectives. First, it analyzes the grammatical and semantic features of institution names and proposes institution name combination rules. It then constructs an institution name normalization framework model, proposes an entity boundary identification method for institution names, compiles a multi-level institution vocabulary, and proposes an institution name normalization method. Finally, empirical testing validates the model.

2 Related Research

Domestic and foreign scholars have proposed various methods and strategies for institution name standardization from the perspective of institution identification, mainly including: (1) Rule-based methods. Shen Jiayi et al. proposed a rule-based Chinese institution identification method for web text data, using an institution suffix lexicon, rule matching, and Bayesian models to identify right and left boundaries. Yang Bo et al. proposed a rule- and statistics-based institution name mapping algorithm for WOS bibliographic data. (2) Statistics-based methods. Hu Wanting et al. proposed an institution name identification method based on word frequency statistics using Baidu Baike entries. Mahemuti et al. proposed a Uyghur institution name identification method based on conditional random field models. Yang Ruixian and Mao Yilei proposed a scientific research institution naming identification method combining rules and vector space models. (3) Chinese institution name normalization. Jia Junzhi et al. constructed an institution name feature vocabulary using the CNKI database and proposed a Chinese institution name normalization method based on TF-IDF and K-means clustering algorithms. Yang Yihong et al. constructed a Chinese multi-level institution vocabulary using the Wanfang database and the knowledge organization method of thesauri. Zeng Jianxun and Jia Junzhi introduced Schema vocabulary to construct a semantic model for institution name standardization data. Sun Haixia et al. used the Chinese biomedical literature database as a corpus and proposed a Chinese institution name normalization method based on K-means clustering algorithms.

3 Institution Name Normalization Framework Model and Algorithm Introduction

[Figure 1: see original paper] Institution Name Normalization Framework Model

3.1 Institution Name Data Collection and Preprocessing

Data collection and preprocessing includes three steps: data source selection, special punctuation preprocessing, and institution field extraction. (1) Data

source selection. Chinese institution name data mainly comes from VIP, CNKI, and Wanfang databases. This study selects VIP as the data source. (2) Special punctuation preprocessing. Due to non-standard institution cataloging, special characters such as spaces, the Chinese enumeration comma (、), forward slash (/), parentheses (()), etc. are used as institution boundary identifiers, which need to be replaced with semicolons during preprocessing. (3) Institution field extraction. In VIP's institution field, the left boundary delimiter is generally "]", the right boundary delimiter is ";", and the middle delimiter is ",", containing one or more levels of institution full names, country, city, postal code, address, etc. After institution field preprocessing, an institution field dataset is obtained, as shown in Table 1 .

3.2 Institution Name Entity Boundary Identification

The general expression for institution names is: F + M* + S. Using the NLPIR tool for word segmentation annotation and word frequency statistics, institution ending identifiers can be obtained. Taking universities as an example, institution name combination rules and ending identifiers are shown in Table 2 .

The general expression for institution level combination rules is: First-level institution + [Second-level institution] + [Third-level institution] + … + [N-level institution]. Based on the institution field dataset, using SQL queries containing Chinese terms for university (大学), college (学院), and department (系) to count the co-occurrence frequency of institution ending identifiers at double and triple levels, generating level combination rules. For universities, institution level combination rules and ending identifiers include: (1) University + [College/Department] + [Department/Office] + [Laboratory/Research Institute/Design Institute/Research Center]; (2) University + [College/Department] + [Department/Office]; (3) University + [College/Department]; (4) University + [Department/Office]; (5) University + [College/Department] + [Laboratory/Research Institute/Design Institute/Research Center]; (6) University + [Department/Office] + [Laboratory/Research Institute/Design Institute/Research Center]; (7) University + [Laboratory/Research Institute/Design Institute/Research Center]; (8) [College/School] + [Department/Office] + [Laboratory/Research Institute/Design Institute/Research Center]; (9) [College/School] + [Department/Office]; (10) [College/School] + [Laboratory/Research Institute/Design Institute/Research Center]; (11) College + [College/Laboratory/Science Park/Co., Ltd./State-owned Assets Office/Selection and Training Office/Office/Kindergarten/Hospital/Team/Station/Institute/Factory].

This paper combines word segmentation and named entity boundary identification to propose an institution name entity boundary identification algorithm. Input: institution field dataset and institution level combination rules. Output: institution name dataset. Process: (1) First-level institution identification. Using a block algorithm based on equivalent matching, institution names, country, address, city, and postal code in entity attributes are defined as five block keys

to construct data record filtering conditions for grouping the institution field dataset, achieving first-level institution identification. (2) Part-of-speech tagging. Using the NLPIR tool with the Chinese Academy of Sciences secondary annotation set for POS tagging. Results show tokens such as '上海' (Shanghai) tagged as/ns, '交通' (Jiao Tong) as/n, '大学' (University) as/n, '电子' (Electronic) as/n, '信息' (Information) as/n, '与' (and) as/cc, '电气' (Electrical) as/n, '工程' (Engineering) as/n, '学院' (College) as/n, '自动化' (Automation) as/vn, '系' (Department) as/n, '系统' (System) as/n, '控制' (Control) as/vn, '与' (and) as/cc, '信息' (Information) as/n, '处理' (Processing) as/vn, '教育部' (Ministry of Education) as/nt, '重点' (Key) as/n, '实验室' (Laboratory) as/n. (3) Determine institution suffix words. Based on POS tagging results, find all institution suffix words in the institution field. From the tokens for department (系) tagged as/v and system (系统) tagged as/n, the term for department is determined as an atomic institution suffix word, while system is an attributive modifier. Finally, four institution suffix words are found: university (大学), college (学院), department (系), laboratory (实验室), and one attributive modifier: system (系统). (4) Determine the right boundary of institution full names. Match institution level combination rules and use "#" as a delimiter to mark the right boundary of each level institution name. The example result after right boundary marking: "Shanghai Jiao Tong University#School of Electronic Information and Electrical Engineering#Department of Automation#Key Laboratory of System Control and Information Processing of Ministry of Education#".

3.3 Compilation of Chinese Multi-level Institution Vocabulary

The basic steps for compiling Chinese multi-level institution vocabulary are: (1) Manually collect standard names to create a basic institution standard name vocabulary. Based on department settings, institutional settings, and historical evolution columns on first-level institution homepages, manually organize standard names of first-level and below-second-level institutions. Baidu Baike, institution establishment news, and institution code tables can also be used. (2) Identify hierarchical relationships to generate a Chinese multi-level vocabulary to be merged. Propose a block algorithm based on co-occurrence relationships and equivalent matching. Based on institution level combination rules, calculate double and triple co-occurrence frequencies, extract co-occurrence relationships between institution entities by setting co-occurrence frequency thresholds, and determine hierarchical relationships among first-level, second-level, and third-level institutions. Generate a Chinese multi-level vocabulary that has not identified identity relationships. (3) Identify identity relationships to generate an uncoded Chinese multi-level vocabulary. Use an edit distance-based similarity algorithm to identify identity relationships, merge textually similar names, and generate an uncoded Chinese multi-level vocabulary. The basic vocabulary can be added as seeds to clustering to improve efficiency and quality. (4) Identify successive relationships. For first-level institutions, manually organize and determine their successive relationships, mainly based on renaming documents released by the Ministry of Education and first-level institution homepages,

institution establishment news, and encyclopedia entries. For second-level institutions, based on identified hierarchical and identity relationships, considering journal publication cycles, use a final appearance year gap greater than 2 as the temporal division standard, and use the number of third-level institutions affiliated with two second-level institutions greater than 3 or overlap exceeding 60% as similarity criteria. According to these two standards, determine their successive relationships. The main observation is: if a college is renamed but not all its subordinate departments, institutes, and laboratories are renamed, the unchanged departments, institutes, and laboratories can identify two institutions with successive relationships but dissimilar text. Since third-level institutions appear with low frequency in bibliographic data and require author information for judgment, this paper treats successive relationships of third-level institutions as identity relationships. (5) Assign unique identifiers to institutions to generate Chinese multi-level institution vocabulary. Use coding to compile a Chinese multi-level institution vocabulary based on unique identifiers.

The significance of compiling Chinese multi-level institution vocabulary includes: (1) The vocabulary serves as a mapping between bibliographic data and institution name standardization data, and vocabulary generation is a data cleaning process of deduplication and merging operations based on co-occurrence matrices. (2) Through Chinese multi-level institution vocabulary, mapping relationships between various variant names and standard names are established to achieve institution name standardization. (3) The automatic vocabulary compilation method ensures institution identification accuracy while saving manpower, providing a new method for institution name standardization construction. (4) Applying Chinese multi-level institution vocabulary identifies institution names in massive data through exact matching to achieve institution name normalization.

3.3.1 Block Algorithm Based on Co-occurrence Relationships and Equivalent Matching This paper draws on linked data ideas to propose a block algorithm based on co-occurrence relationships and equivalent matching. Input: institution name dataset. Output: institution name data block results. Process: (1) Establish co-occurrence matrix. Based on institution name entity boundary identification, establish two-dimensional co-occurrence matrices (University-College) and three-dimensional co-occurrence matrices (University-College-Department/Institute/Laboratory) for each level institution. (2) Set co-occurrence frequency threshold. Extract co-occurrence relationships between institution entities by setting thresholds. Use co-occurrence matrices to simultaneously extract hierarchical and co-occurrence relationships between institution entities, revealing semantic relationships through hierarchical relationships to solve the problem of lacking semantic relationship definitions between various institution name types in traditional rule-based methods, compensating for defects of methods relying solely on formal name matching. (3) Define block keys. Use each level institution (University, College, Department) as entity attributes, define one or more block keys, and map them to different data blocks based on

key values in institution name data to improve matching efficiency.

3.3.2 Edit Distance-Based Similarity Algorithm On the basis of data blocking, sort author affiliation fields and adopt a sliding window method with a width of 30 and step size of 1. Use the Levenshtein distance, Jaro distance, and Jaro-Winkler distance to measure string similarity, set similarity thresholds, and complete institution name merging. When the number of prefix characters is greater than or equal to 1, the Jaro-Winkler distance adjusts based on the Jaro distance to characterize string similarity of the same prefix part. The Jaro distance algorithm is:

$$Jaro(str1, str2) = \frac{c - t/2}{|str1|} + \frac{c - t/2}{|str2|}$$

where $|str1|$ and $|str2|$ are string lengths; c is the number of common characters between two strings, requiring $str1[i] = str2[j]$ and $|i - j| \leq \frac{\min\{|str1|,|str2|\}}{2}$; t is the number of transpositions, counting mismatches when comparing the i-th common character of two strings.

The Jaro-Winkler distance algorithm is:

$$d_w = d_j + [lp(1 - d_j)]$$

where d_j is the Jaro distance between two strings; l is the number of identical prefix characters, with a maximum value of 4; p is a constant, with a maximum of 0.25, which Winkler set to 0.1.

3.3.3 Multi-level Vocabulary Compilation Based on Unique Identifiers First-level institution unique identifiers include national organization unified social credit codes, institution codes in the national list of regular higher education institutions, etc. Below-second-level institution unique identifiers include internal institution codes, database field codes, etc. Additionally, email, postal code, address, institution URL, etc., which are unique and have mapping relationships with institution entities, can be regarded as institution unique identifiers. This paper uses institution codes from the national list of regular higher education institutions as first-level institution codes, five-digit strings as below-second-level institution codes, and supplementary codes for collaborative innovation centers, research institutes, research centers, bases, laboratories, etc., co-built by universities and external institutions that are not affiliated with other internal second-level institutions. “UC:”, “SC:”, and “OC:” represent first-level institution codes, below-second-level institution codes, and supplementary codes, respectively.

An example of Chinese multi-level institution vocabulary is shown in Table 3 . From Table 3, Nanyang Public School and Shanghai Jiao Tong University have

a successive relationship; the National Defense Science and Technology Key Laboratory of Micro/Nano Manufacturing Technology and the School of Electronic Information and Electrical Engineering have a hierarchical relationship; the Department of Micro-Nano Electronics and the Department of Micro-Nano Electronics have an identity relationship.

3.4 Institution Name Normalization

To achieve model generalization, this paper proposes an exact matching-based institution name normalization algorithm. Input: institution name dataset, multi-level institution vocabulary. Output: institution name normalization results. Process: (1) Load Chinese multi-level institution vocabulary. When entity boundary-identified institution data exactly matches the Chinese multi-level institution vocabulary, annotate its institution code. (2) When exact matching fails, perform word segmentation annotation on both the entity boundary-identified institution data and institution names in the Chinese multi-level institution vocabulary, remove all punctuation marks, and perform exact matching again. After successful matching, annotate its institution code.

When exact matching fails, for example, the word segmentation result of institution name "School of Naval Architecture, Ocean and Civil Engineering" is "Naval Architecture/n, /wn Ocean/n and/cc Architecture/vn Engineering/n College/n". After removing symbols, the variant name transforms into "School of Naval Architecture Ocean and Civil Engineering", which successfully matches the standard name exactly.

4 Experiments and Results Analysis

4.1 Data Collection and Preprocessing

To verify model effectiveness, the VIP database was selected as the data source. Using Shanghai Jiao Tong University as the target institution, the search query "S=(Shanghai Jiao Tong University OR Shanghai Jiaotong OR 15 affiliated hospitals)" was used with a time span of 2008-2018. A total of 145,538 documents were retrieved on March 21, 2019. After data preprocessing, 233,998 institution name data entries were obtained, with 121,065 journal articles published with Shanghai Jiao Tong University as the first institution.

To evaluate model applicability to different institution types, samples of university-named and college-named institutions were used to test the research method's effectiveness. Changshu Institute of Technology was further selected as a college-named institution for retrieval, yielding 8,292 documents from 2008-2018.

4.2 Institution Name Entity Boundary Identification

Using the NLPIR tool for word segmentation annotation and based on institution level combination rules, entity boundary identification was performed

on institution names. Partial results for Shanghai Jiao Tong University name entity boundary identification are shown in Table 4 .

4.3 Compilation of Chinese Multi-level Institution Vocabulary

4.3.1 Block Algorithm Based on Co-occurrence Matrix and Equivalent Matching Using three-dimensional co-occurrence matrices of multi-level institutions, co-occurrence frequencies of first-level, second-level, and third-level institutions were calculated. Taking the School of Naval Architecture, Ocean and Civil Engineering as an example, the co-occurrence frequency threshold can be set to 3. Partial calculation results of multi-level institution co-occurrence relationships are shown in Table 5 . Using the block algorithm based on equivalent matching for data blocking, when using second-level institutions as block keys, various variant names of the same institution entity can be discovered through similarity calculations in the data block with key value "School of Naval Architecture Ocean and Civil Engineering", such as "Department of Civil Engineering", "Civil Engineering Department", "Department of Architecture", "Architecture Department", etc. When using third-level institutions as block keys, various variant names of the same institution entity, such as abbreviations and full names like "School of Naval Architecture and Civil Engineering" and "School of Naval Architecture Ocean and Civil Engineering", can be discovered in the data block with key value "Department of Engineering Mechanics".

4.3.2 Edit Distance-Based Similarity Algorithm On the basis of data blocking, the sliding window method is adopted with a width of 30 and step size of 1. Levenshtein distance, Jaro distance, and Jaro-Winkler distance are used to measure string similarity, with similarity thresholds set to complete institution name merging. When the number of prefix characters is greater than or equal to 1, Jaro-Winkler distance adjusts on the basis of Jaro distance to characterize string similarity of identical prefix parts. String similarity calculation results based on edit distance are shown in Table 6 . Setting the Jaro-Winkler distance threshold to 0.75 can obtain string pairs with high similarity as initial data for manual verification. Finally, the compiled university institution multi-level vocabulary contains 3,528 institution names, including 2,834 variant names and 694 standard names. There are 160 second-level institution standard names, including 66 co-construction platforms between the university and other institutions; 478 third-level institution standard names; and 56 fourth-level institution standard names, mainly research institutes and departments (such as the Economics Department of the Economics School of the Antai College of Economics and Management).

4.4 Institution Name Normalization

By randomly selecting 1,000 institution normalization result data entries from 233,998 Shanghai Jiao Tong University data entries for manual verification, the final institution name normalization experimental results are shown in Ta-

ble 7 . After manual verification, there are 8 errors in institution name entity boundary identification, including "Center#Hospital#" and "Medical College#Hospital#" due to failure to merge adjacent institution identifier words. There are 5 data entries where Shanghai Jiao Tong University first-level institution was not identified, with 1 first-level institution identification error. There are 2 second-level institution identification errors for Shanghai Jiao Tong University, caused by entity boundary identification errors and multi-level vocabulary errors. Experimental results show that institution name entity boundary identification accuracy is 99.2%; Shanghai Jiao Tong University first-level institution identification accuracy is 99.9%, recall is 99.3%, and F-measure is 99.6%; second-level institution identification accuracy is 99.7%, recall is 95.5%, and F-measure is 97.6%. Among 31 data entries where second-level institutions were not identified, 6 were due to variant names not included in the multi-level institution vocabulary, such as "Shanghai Ninth People's Hospital" and "Jiulong Hospital of Suzhou, Jiangsu Province"; 5 were due to first-level institutions not being identified; the remaining 20 data entries only bear Shanghai Jiao Tong University without second-level or lower institution information, which cannot be identified using only author affiliation fields in bibliographic data and require author information.

By randomly selecting 1,000 institution normalization result data entries from 11,569 Changshu Institute of Technology data entries for manual verification, experimental results show that first-level institution identification accuracy is 100%, recall is 100%, and F-measure is 100%; second-level institution identification accuracy is 99.8%, recall is 80.5%, and F-measure is 89.1%. Among 148 data entries where second-level institutions were not identified, all only bear Changshu Institute of Technology without second-level institution information.

5 Discussion

The era of big data and academic big data's new characteristics call for innovation in institution identification models. Examining institution identification from a linguistic perspective, considering three identification dimensions of identity, successive, and hierarchical relationships, and integrating three levels of model input, process, and output, combining rule- and statistics-based automatic identification with manual verification to improve data reliability in scientific research management, discipline evaluation, and subject services under big data environments, constructing a data-driven institution name normalization model, and reconstructing the traditional two-stage model of automatic assignment + manual assignment based on data cleaning platforms represents a scientific strategy for advancing institution identification in the new era.

The constructed institution name normalization model mainly explores input data, normalization process, and institution vocabulary compilation, specifically reflected in: First, input data quality control. The model uses author affiliation fields as the sole data source, supplemented by rules and authoritative, accurate institution code tables, ensuring data reliability at the model

input end and avoiding data pollution caused by ambiguous authors and non-standard postal codes when referencing author and postal code information. Second, white-box model. The model uses exact matching for institution identification, overcoming the black-box limitations of machine learning algorithms and the difficulty of correction. The core and foundation of exact matching is the compiled Chinese multi-level institution vocabulary, whose production combines automatic identification and manual verification. The number of variant names in a dataset is limited and can be completely manually verified, making the vocabulary represent the identification level of subject librarians and ensuring vocabulary accuracy. Especially when identification error samples are found during manual verification or practical application, batch correction can be efficiently completed by modifying the vocabulary. Third, multi-level institution vocabulary compilation based on entity relationship identification. The paper proposes a block algorithm based on co-occurrence relationships and equivalent matching, and uses an edit distance-based similarity algorithm to automatically identify hierarchical and identity relationships between institution entities, reducing manual costs in multi-level institution vocabulary production and avoiding the shortcomings of keyword frequency-based methods that suffer from interference between multi-level institutions. For identification and updating of successive relationships, regular maintenance of multi-level institution vocabulary can be performed through multiple channels such as institution change reports, periodically released internal institution code tables, and bibliographic data extraction, reducing the time lag impact of successive relationship identification based solely on bibliographic data extraction.

Empirical results demonstrate that the institution name normalization model can achieve entity boundary identification and second-level institution identification for both university-named and college-named institution types, verifying model effectiveness and providing insights for normalizing other institution name types. However, some issues remain to be addressed, such as empirical verification when model output assigns literature to institutions below the third level, automatic identification of successive relationships, and validating model effectiveness on other bibliographic datasets.

References

[1] Jia Junzhi, Zeng Jianxun, Li Jiejia, et al. Implementation of scientific research institution name normalization[J]. Library and Information Service, 2018, 62(13): 103-110.

[2] Zeng Jianxun, Wang Lixue. Construction method of normative documents for knowledge evaluation[J]. Library and Information Service, 2012, 56(10): 101-106.

[3] Zeng Jianxun, Jia Junzhi. Construction of semantic model for institution name standardization data[J]. Journal of Academic Libraries, 2019, 37(1): 42-47.

[4] Liu Bing. Sentiment Analysis: Mining Opinions, Sentiments, and Emotions[M]. Translated by Liu Kang, Zhao Jun. Beijing: Mechanical Industry Press, 2017(7): 1.

[5] Shen Jiayi, Li Fang, Xu Feiyu, et al. Identification of Chinese organization names and abbreviations[J]. Journal of Chinese Information Processing, 2007(6): 17-21.

[6] Yang Bo, Yang Junwei, Yan Sulan. Research on institution name standardization based on rules[J]. New Technology of Library and Information Service, 2015(6): 57-63.

[7] Hu Wanting, Yang Yan, Yin Hongfeng, et al. An organization name identification method based on word frequency statistics[J]. Computer Applications Research, 2013, 30(7): 2014-2016.

[8] Mahemuti Maimaiti, Wang Lulu, Tuergen Yibulayin, et al. Uyghur institution name identification based on conditional random fields[J]. Computer Engineering and Design, 2019, 40(1): 273-278.

[9] Yang Ruixian, Mao Yilei. Research on naming identification method of Chinese scientific research institutions for knowledge evaluation[J]. Journal of Intelligence Science, 2015, 34(7): 179-183.

[10] Yang Yihong, Li Yaping, Zhang Lili, et al. Compilation of multi-level institution vocabulary and its application in bibliometric evaluation and scientific research performance management[J]. Digital Library Forum, 2013(6): 57-63.

[11] Sun Haixia, Li Junlian, Wu Yingjie. Research on institution normalization based on K-means[J]. Journal of Medical Informatics, 2013, 34(7): 41-44+71.

[12] Shen Derong, Kou Yue, Nie Tiezheng, et al. Entity Recognition Technology[M]. Beijing: Mechanical Industry Press, 2017(9): 45-50.

[13] JARO M A. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida[J]. Journal of the American Statistical Association, 1989, 84(406): 414-420.

[14] WINKLER W E. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage[C]//Proceedings of the section on survey research methods, Washington, DC: American Statistical Association, 1990: 354-359.

Author Contributions: Yang Zhao: Responsible for topic selection, research framework design, and paper writing; Ren Juan: Responsible for data collection and organization, and paper writing.

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

Source: ChinaXiv — Machine translation. Verify with original.