Implementation of Research Institution Name Normalization for Postprints

Authors: Jia Junzhi, Zeng Jianxun, Li Jiejia, Fu Xiaomei

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

Abstract

[Purpose/Significance] Institution names are numerous and heterogeneous. Institution name normalization can aggregate the canonical name of the same institution along with its non-canonical names from different periods and in various forms, thereby improving both the recall and precision of search and retrieval, and facilitating interoperability with other systems as well as resource sharing. [Method/Process] Based on an analysis of institution name string characteristics and the K-means algorithm, the edit distance algorithm is employed to achieve preliminary clustering of first-level institution names. Subsequently, leveraging the preliminary clustering results, the TF-IDF algorithm is utilized to calculate the weight of each term in institution names, thereby clustering institution names around cluster centers via the K-means algorithm and assigning a unique identifier to the institution names within each cluster. [Results/Conclusion] This method can achieve normalization of different forms of canonical names for the same institutional entity and improve the accuracy of institution name clustering; however, the selection of K values and distance measurement methods remains to be optimized.

Full Text

Preamble

ChinaXiv Collaborative Journal, Vol. 62, No. 13, July 2018

Realization of Research Institution Name Normalization

Jia Junzhi[1], Zeng Jianxun[2], Li Jiejia[1], Fu Xiaomei[1]

[1] School of Economics and Management, Shanxi University, Taiyuan 030006
[2] Information Resources Center, Institute of Scientific and Technical Information of China, Beijing 100038

Abstract

[**Purpose/Significance**] Institution names are numerous and complex. Institution name normalization aggregates the authoritative name of an institution together with its non-standard variants from different time periods and expressions, thereby improving both the recall and precision of information retrieval. It also facilitates interoperability with other systems and enables resource sharing.

[**Method/Process**] Based on an analysis of institution name characteristics and the K-means algorithm, this study employs edit distance similarity algorithm to achieve preliminary clustering of first-level institution names. It then utilizes the initial clustering results to calculate term weights using TF-IDF, enabling institution names to cluster around central points via K-means algorithm, with each cluster receiving a unique identifier.

[**Result/Conclusion**] This method successfully normalizes different forms of names belonging to the same institution entity and improves clustering accuracy. However, optimization is still needed for selecting the K value and distance measurement methods.

Keywords: Research institution name; Clustering; K-means

1. Introduction

Research institutions are entities such as universities and research institutes engaged in scientific research. Due to diverse naming conventions, frequent mergers, changes, and renaming, as well as unclear relationships between institutions, identifying institution names presents significant challenges. Furthermore, researchers publishing at different times often use different expressions for the same affiliation, including authoritative names, former names, abbreviations, and erroneous recordings. This variability affects information retrieval, statistical analysis, and bibliometric evaluation based on institution names. Institution name normalization aims to consolidate different expressions of the same institution name, establish correspondences between authoritative and variant names, and achieve institution identification by assigning unique identifiers.

For institution name authority control, foreign practices are relatively mature, managing institution aliases and multilingual names through authority files and databases. The Virtual International Authority File (VIAF), initiated in 2003 at the IFLA conference in Berlin by the Library of Congress, German National Library, and OCLC, links authority files from various libraries to connect different name forms for the same person or organization. Linking and Exploring Authority Files (LEAF), launched in 2001 by the European Commission, aims

to develop a distributed retrieval system architecture linking dispersed records of different name forms to their authority records.

In China, name authority control started later. In the 1990s, the National Library of China began Chinese name authority control. In 2009, the National Library of China, JULAC-HKCAN, CSS, and CALIS jointly established CNAJDS (China Name Authority Joint Database Search System), integrating name authority data from member institutions to achieve resource sharing. Additionally, the Chinese Academy of Sciences built its institution name authority database to comprehensively construct normalized descriptions of CAS institutions, enabling rapid registration of authoritative and alias names, 梳理 institutional relationships, and providing institution name authority services. However, discrepancies exist between different databases, with the same institution using inconsistent names and lacking automatic mechanisms to identify renaming and alias relationships. As institution name quantities increase, manual processing becomes increasingly complex, making clustering-based identification a viable auxiliary solution.

Current research on institution name identification can be divided into rule-based methods and statistical methods. Rule-based approaches use feature word triggers, analyzing grammatical properties, semantic characteristics, and organizational patterns to summarize rules and patterns for identification. Statistical methods train large-scale corpora to build models, including chunk analysis, decision trees, conditional random fields, support vector machines, and hidden Markov models.

For name normalization, Y. Jiang et al. used normalized compression distance for clustering different names of the same institution. Yang Yihong et al. compiled multi-level institution vocabularies to solve normalization problems in massive data for bibliometric evaluation. Sun Haixia et al. employed K-means algorithm with frequent word sets for institution name normalization, though center selection needed improvement. Xian Xin used K-nearest neighbor combined with edit distance similarity for normalization but addressed renaming relationships minimally.

Existing institution name normalization has achieved clustering of aliases and authoritative names but relies mainly on frequency statistics without organically combining name identification and normalization. During clustering, only single methods are applied after center determination, with limited consideration of renaming relationships. This study builds institution feature word tables through term analysis, identifies hierarchical structures (first-level and second-level institutions), and applies edit distance algorithm, TF-IDF method, and K-means algorithm in a two-stage approach to optimize clustering effects.

---

2. Building Institution Feature Word Tables

Feature word tables are constructed to effectively identify first-level and second-level institution names, typically comprising terms indicating institutional nature or type. Analysis reveals that institution names generally follow an "A+B" pattern as a naming phrase centered on B. Component A consists of verbs, locatives, ordinals, adjectives, etc., with variable length, while component B is relatively fixed and limited to terms like "University," "College," "Research Institute," etc. Thus, feature word tables can identify the B component, while A can be analyzed through part-of-speech tagging and sequence pattern matching.

2.1 Data Sources

Institution name normalization requires capturing evolution patterns. This study collects sample data from CNKI. Direct retrieval by institution name affects recall and precision: exact matching ignores literature under other names of the same institution, while fuzzy matching yields overly broad results. Therefore, selected data must include not only non-standard forms like abbreviations and aliases but also reflect temporal evolution.

Considering these factors, this study uses academic journals as the carrier, selecting data from *Library and Information Service*, *Chinese Journal of Computers*, and *Journal of Mechanical Engineering* from 2006-2016 (11 years). The data includes author affiliations, authors, journal names, and publication years [Figure 1: see original paper]. "Author affiliation" data is extracted to explore institution name evolution, yielding 13,839 records. After splitting multi-author records and removing duplicates and noise data (e.g., "AL10", "9AB", "VA"), 6,503 valid records remain.

2.2 Part-of-Speech Feature Analysis of Institution Names

Part-of-speech tagging is a preprocessing step in natural language processing that reveals combination patterns between institution name terms and adjacent components. Using the NLPIR Chinese word segmentation system, 6,503 institution names are segmented into 41,439 words, primarily nouns, verbs, and adjectives [Figure 2: see original paper]. In the tagging set, n=noun, ns=place name, m=numeral, vn=nominal verb, cc=coordinating conjunction, b=distinctive word. Statistics show nouns dominate institution names. Combination sequences fall into four types: Noun + Noun (e.g., "China/ns People/n University/n News/n College/n"); Verb + Noun (e.g., "Wuhan/ns University/n News/n and/cc Communication/vn College/n"); Adjective + Noun (e.g., "Shandong/ns University/n Public/b Health/an College/n"); Ordinal + Noun (e.g., "Guangdong Province/ns Lingnan/ns Business/n First/m Technician/n College/n"). This analysis primarily uses nouns for feature word identification, with sequence patterns providing auxiliary judgment.

2.3 Feature Word Table Determination

In addition to part-of-speech distribution, word frequency statistics are analyzed. Among 82,946 total words, "University" appears most frequently (3,411 times), followed by "College" (2,727 times), indicating they are likely feature words. "University" typically marks first-level institutions, while "College" can indicate either first-level (e.g., "Taiyuan Normal College") or second-level institutions (e.g., "School of Geography, Taiyuan Normal College"). High-frequency words like "Engineering," "Information," and "Science" are mostly industry-direction or specialty-type terms closely related to the original data source.

Feature word selection requires high-frequency, general terms. The threshold between high and low frequency words uses Donohue's formula derived from Zipf's law:

$$T = \frac{-1 + \sqrt{1 + 8 \times I_1}}{2}$$

where I_1 is the number of keywords with frequency 1, and T is the minimum frequency value among high-frequency words (the critical threshold). In this study, $I_1 = 1,090$, yielding $T = 46$. Thus, the high-low frequency threshold is set at 46 occurrences. Words appearing more than 46 times are selected as high-frequency terms and manually reviewed. Eleven feature words are ultimately identified .

---

3. Institution Name Hierarchy Recognition

Before normalization, first-level and second-level institutions must be identified. Word segmentation alone cannot distinguish these hierarchies. Using the constructed feature word table, each institution name string is traversed left-to-right. Symbols are assigned to first-level, second-level, etc., institutions upon successful matching with the feature word table, enabling hierarchical differentiation.

Processing flow: After data preprocessing, institution data is precisely matched against the feature word table. For example, "Shanxi University School of Economics and Management" is annotated as "Shanxi University # School of Economics and Management #", treating "Shanxi University" as first-level and "School of Economics and Management" as second-level. For names with multiple feature words like "Beijing Normal University School of Management Department of Information Management and Information Systems", annotation yields "Beijing Normal University # School of Management # Department of Information Management and Information Systems #". Special cases like "Shanxi University Business College" (a first-level institution) are annotated as "Shanxi University # Business College #", which incorrectly splits it; such issues are addressed in Section 4. Sample results are shown in .

4. Institution Name Normalization

Institution name normalization consolidates different expressions of the same institution name, establishing correspondences between authoritative and variant names through unique identifiers.

4.1 Preliminary Clustering

Edit distance algorithm (Levenshtein distance), proposed by V. I. Levenshtein in 1966, measures the minimum number of edit operations (insertion, deletion, substitution) required to transform a source string into a target string. Similarity (Sim) is calculated as:

$$Sim = 1 - \frac{\text{edit count}}{\text{maximum value}}$$

When no edits are needed (edit distance = 0), similarity is 100% (identical strings). When every character must be edited, similarity is 0% (no similarity). A threshold Y between 0% and 100% determines whether clustering occurs.

First-level institution names are extracted as the sample. The algorithm proceeds as follows: Set similarity threshold Y, use the first record as the initial cluster center, then calculate similarity with remaining records. Records with similarity $\geq Y$ join the cluster. From records with similarity $< Y$, the next unclustered record becomes a new center, repeating until all data is processed [Figure 3: see original paper].

4.2 Normalization Based on TF-IDF and K-means

After preliminary clustering of first-level institution names, TF-IDF values are calculated for terms within each cluster. K-means algorithm then binds first- and second-level institution names for iterative clustering within each first-level cluster, narrowing the scope and achieving final normalization.

TF-IDF (term frequency-inverse document frequency) assesses term importance in a document collection. TF is term frequency in a document; IDF = $\log_2(\text{N/DF})$, where DF is the number of documents containing the term. The core idea: terms frequent in a document but rare across the collection are most discriminative. The algorithm: segment documents, remove stop words, count term frequencies, then compute TF-IDF values.

Traditional K-means predefines cluster centers and iteratively updates them to minimize the objective function. Here, TF-IDF values are embedded into K-means: randomly select initial centers, compute TF-IDF values, calculate distances between remaining data and centers using Euclidean distance:

$$D(X_i, Y_i) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$

Assign data to nearest clusters, recompute centers as geometric means, and iterate until convergence. After clustering, each cluster receives a unique identifier ID [Figure 4: see original paper].

---

5. Extraction of Institution Renaming Relationships

Clustering identifies aliases with high similarity. However, renaming relationships (e.g., "North University of China" formerly "North China Institute of Technology") lack lexical similarity and cannot be summarized through part-of-speech rules. Therefore, author co-occurrence rates are used to mine institutional relationships. Co-occurrence rate is defined as the ratio of common authors (author intersection) to total authors (author union) between two institutions.

After clustering, each institution cluster has a corresponding author set. Renaming relationships are extracted by computing co-occurrence between author sets of different clusters. A threshold X is preset; co-occurrence rates $\geq X$ indicate renaming relationships. The IDs of related clusters are merged (using the more recent name's ID), with older names marked as former names [Figure 5: see original paper].

---

6. Experimental Analysis

The experiment uses MyEclipse with Java for hierarchy recognition, edit distance clustering, and TF-IDF/K-means clustering, with C language for extracting renaming relationships.

6.1 Institution Name Hierarchy Recognition

Using preprocessed data, recognition results for Peking University-affiliated institutions are shown in [Figure 6: see original paper].

6.2 Institution Name Normalization

6.2.1 Preliminary Clustering via Edit Distance First-level institutions are extracted, sorted, and deduplicated, yielding 2,465 records. Edit distance clustering uses an 85% similarity threshold. Results for Hebei institutions are shown in [Figure 7: see original paper]. Special cases include:

1. **Same institution, different recording**: e.g., "Guangdong University of Foreign Studies" vs. "Guangzhou University of Foreign Studies"; "National

University of Defense Technology" vs. "National University of Defense Science and Technology"; "Shanxi Tourism Vocational College" vs. "Shanxi Province Tourism Vocational College."

2. **Different institutions with high similarity**: e.g., locational prefixes ("Beihang University" vs. "Nanjing University of Aeronautics and Astronautics"), modifiers ("Harbin Institute of Technology," "Harbin Engineering University," "Harbin University of Commerce"), or unified naming of affiliated institutions ("Sixth Research Institute of China Aerospace Science & Industry Corp" vs. "Third Research Institute…").

For similar names with comparable frequencies, TF-IDF calculation provides high discrimination, allowing K-means to separate them into distinct clusters. For similar names with large frequency differences, author intersection information is used. The co-occurrence rate from renaming extraction is applied; if it meets the threshold, they are treated as the same institution; otherwise, manual judgment guides separation before clustering.

6.2.2 Clustering Based on TF-IDF and K-means Within each first-level cluster, TF-IDF values are computed for second-level institution terms. K-means binds first- and second-level names for clustering. Results for "Shanxi University" are shown in [Figure 8: see original paper] and . Publication dates are extracted and sorted to establish alias relationships, with the most recent name as the center and others as aliases (e.g., ID "0300101" centers on "Shanxi University School of Economics and Management").

Some clustering errors occur, such as conflating "Shanxi University School of Computer and Information Technology" with "Shanxi University School of Economics and Management" due to high frequency of shared terms like "information," preventing 100% accuracy.

6.3 Extraction of Institution Renaming Relationships

After integrating author data, a 15% co-occurrence threshold is set. C language implementation extracts renaming relationships [Figure 9: see original paper]. For "North China Institute of Technology" and "North University of China," the co-occurrence rate reaches 15%, confirming a renaming relationship. The more recent name's ID is assigned to the merged cluster.

6.4 Experimental Evaluation

Clustering effectiveness is evaluated through precision and recall:

$$\text{Precision: } P = \frac{A}{A+B}$$

$$\text{Recall: } R = \frac{A}{A+C}$$

where A = correctly clustered names, B = incorrectly clustered names, C = unclustered but correct names. Manual verification shows precision $ $80% and recall $ $75% for second-level clustering under first-level institutions, indicating good performance with room for algorithmic optimization.

As science and technology advance, research talent emerges continuously, and institutions evolve through recording omissions, mergers, and personnel changes, creating diverse name expressions. This study transforms the direct K-means approach by first identifying hierarchies via feature word tables, then applying edit distance for preliminary first-level clustering to narrow scope and improve efficiency. TF-IDF enhances precision and recall. This two-stage clustering optimizes results, 挖掘 relationships between names, establishes rules for quantitative research, and reduces manual authority file construction workload.

However, data spans only 11 years, limiting analysis of long-term evolution (e.g., college-to-university upgrades, mergers). Larger temporal datasets are needed. K-value determination still impacts results and requires optimization via machine learning or statistical methods. These issues warrant further investigation.

References

[?] VIAF[EB/OL].[2017-03-06]. http://www.oclc.org/en/viaf.html.
[?] LEAF-linking and exploring authority files[EB/OL].[2018-03-27]. http://irsr.llas.ac.cn/institution/.
[?] Chinese Name Authority Joint Database Search System[EB/OL].[2017-12-01]. http://cnass.cccna.org/jsp/index.jsp.
[?] Chinese Academy of Sciences Institution Name Authority Database[EB/OL].[2017-03-08]. http://irsr.llas.ac.cn/institution/.
[?] Zhang Xiaoheng, Wang Lingling. Recognition and analysis of Chinese institution names[J]. Chinese Information Processing, 1997, 11(4): 21-32.
[?] Shen Jiayi, Li Fang, Xu Feiyu, et al. Recognition of Chinese organization names and abbreviations[J]. Journal of Chinese Information Processing, 2007, 21(6): 17-21.
[?] Chen Xiao, Liu Hui, Chen Yuquan. Chinese organization name recognition based on support vector machines[J]. Computer Applications Research, 2008, 25(2): 362-364.
[?] Yu Hongkui, Zhang Huaping, Liu Qun, et al. Chinese named entity recognition based on hierarchical hidden Markov models[J]. Journal of Communications, 2006, 27(2): 87-94.
[?] Ye Linli, Huang Rima. Chinese institution name recognition combining decision tree methods[J]. Fujian Computer, 2007(12): 184-184.
[?] Yin Jihao, Fan Xiaozhong, Zhao Panchao, et al. Chinese institution name recognition based on chunk analysis technology[J]. Journal of Harbin Engineering University, 2006, 27(S1): 466-.

[?] JIANG Y, ZHENG H T, WANG X, et al. Affiliation disambiguation for constructing semantic digital libraries[J]. Journal of the American Society for Information Science & Technology, 2011, 62(6): 1029-1041.
[?] Yang Yihong, Li Yaping, Zhang Lili, et al. Compilation of multi-level institution vocabularies and application in bibliometric evaluation and research performance management[J]. Digital Library Forum, 2013(6): 57-63.
[?] Sun Haixia, Li Junlian, Wu Yingjie. Research on institution normalization based on K-means[J]. Journal of Medical Informatics, 2013, 34(7): 41-44.
[?] Xian Xin. Research on institution authority document structure and construction methods[D]. Beijing: Institute of Scientific and Technical Information of China, 2015.
[?] DONOHUE J C. Understanding scientific literatures: a bibliometric approach[M]. Cambridge: The MIT Press, 1973: 49-50.
[?] Zhang Chengzhi. A multi-layer feature-based string similarity computation model[J]. Journal of Intelligence, 2005, 24(6): 696-701.
[?] LEVENSHTEIN V I. Binary codes capable of correcting deletions, insertions and reversals[J]. Soviet physics doklady, 1966, 10(1): 707-710.
[?] Wu Jun. The beauty of mathematics[M]. Beijing: Posts & Telecom Press, 2014.
[?] Li Fei, Xue Bin, Huang Yalou. K-means clustering algorithm with optimized initial centers[J]. Computer Science, 2002, 29(7): 94-96.
[?] He Xiaoqun. Multivariate statistical analysis[M]. Beijing: China Renmin University Press, 2012.

Author Contributions:
Jia Junzhi: Research design, framework, and manuscript revision
Zeng Jianxun: Research guidance
Li Jiejia: Data collection and manuscript writing
Fu Xiaomei: Data collection and manuscript writing

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