ELAD: An Entity Linking Based Affiliation Disambiguation Framework

ZHOU SHAO†1,4,5,6, XIANGYING CAO2, SHA YUAN3,4, AND YONGLI WANG1, (Member, IEEE)

1School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
2Information School, Renmin University of China, Beijing 100872, China
3Beijing Academy of Artificial Intelligence, Beijing 100086, China
4Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
5School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang 621010, China
6Graduate School of China Academy of Engineering Physics

Corresponding author: Yongli Wang (yongliwang@mail.njust.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61941113 and Grant 61806111, in part by the Fundamental Research Fund for the Central Universities under Grant 30918015103 and Grant 30918012204, in part by the Nanjing Science and Technology Development Plan Project under Grant 201805036, in part by the “13th Five-Year” equipment field fund under Grant 61403120501, in part by the China Academy of Engineering Consulting Research Project under Grant 2019-ZD-1-02-02, in part by the National Social Science Foundation under Grant 18BTQ073, and in part by the State Grid Technology Project under Grant 5211XT190033.

ABSTRACT The number of papers has exploded as more and more people and more types of institutions participating in scientific research. At the same time, institution name disambiguation (IND) is getting more sophisticated, which is critical for research assessment, scholar alignment, etc. Previous knowledge-based and rule-based methods require knowledge and rules prepared in advance, which cannot cope with growing and changing data and learning rules, especially for data with a long period and abundant sources. This paper proposes an automatic learning framework to solve the problem, which is based on entity linking, entity type recognition, candidate generation, and result selection. Experiments show that precision and recall is much higher than the traditional method, ELAD learns more knowledge from the knowledge graph, and it can deal with ever-changing and ever-increasing data. What’s more, it solves many problems that cannot be solved by traditional methods: the connection between institution entities, mistakes correction, and the reduction of manual and pre-prepared knowledge. At last, for the case study, we develop two applications based on ELAD which proves its reliability.

INDEX TERMS Institution name disambiguation (IND), knowledge graph, entity linking.

I. INTRODUCTION

The primary purpose of IND in academic big data is to identify and map the top-level institutions in affiliations to real-world institutional entities. IND is essential for research assessment [1], institution cooperation network [25], scholar name disambiguation [27], scholar trajectory [18], talent flow [13], management of scientific papers [10], assessment of institutional research performance [17], institution ranking [4], etc. With the progress of modern science and technology, the number of scientific research papers surges. A statistic shows that the average global growth rate has been sustained at about 15% for both papers and patents in recent years [28]. On the one hand, the number of newly published papers is increasing exponentially. On the other hand, the backward data storage technology in the past years has caused a lot of non-traceable problems. What is more, to meet the needs of the rapid development of modern technology, all kinds of joint research laboratories, enterprise research institutes, new specialized research academics, and other new research institutions emerge in an endless stream as well. The historical technical problems, the explosion of paper data, and the emergence of new research institutions make the background of IND problem rather complicated.

At the same time, the causes of the problem are so complex, such as different translation methods, different spelling languages, institutional transition, spelling errors, institution versus divisions [8], different writing styles, typographical or OCR mistakes, abbreviations, translation mistakes [22], omissions [6], various institutions have the same name, etc.
Moreover, it is even hard to tell which causes the institution name disambiguation error for specific affiliations. Although traditional methods can solve problems within a finite set with strict rules and knowledge, the issue becomes complicated to deal with when faced with a large number of growing data sets.

This paper focuses on addressing the IND problem based on entity linking, which will reduce the reliance on pre-defined rules and pre-prepared knowledge. What is more, the method can deal with the growing and changing types of institutions in affiliations with the help of a regularly updated knowledge graph.

II. RELATED WORK

Previous work in solving this problem can be divided into three main categories: rule-based methods, knowledge-based methods, and the combination of these two methods. The main characteristics of the rule-based methods includes: (1) a productive system of rules based on common instances, usually implemented by regular expressions; (2) a rich additional keywords library based on regulations, e.g., city/state/country names, keywords for organizations, etc.; (3) similarity matching algorithm or clustering. Knowledge-based methods usually include the following properties: (1) pre-prepared knowledge bases, e.g., institutions and their introductions; (2) manual handling.

De Bru and Moed [7] formulate the most simple IND rules. NEMO is a typical rule-based system, which multi-layered rule matching with multiple dictionaries [11]. The normalization process involves clustering based on weighted local sequence alignment metrics to address synonymy at the word level, and local learning based on finding connected components to address synonymy [11]. Jiang et al. propose a clustering method based on normalized compression distance to group the affiliations that denote the same institute [10]. Cuxac et al. [6] apply both supervised and semi-supervised approaches for affiliations clustering. Huang et al. propose a rule-based algorithm, of which precision is high, and recall is low [8]. Sun et al. build an affiliation knowledge base named Authority File for Affiliations (AFA) based on ontology principles [20]. And they also established seven name matching rules based on their regions, types, and naming characteristics [21].

Aumueller et al. determine the similarity of affiliations based on how the URLs in the result sets of affiliation web searches overlap [3]. Based on that, they also proposed an approximate string metric that handles acronyms and abbreviations [2]. Morillo et al. propose a new semi-automatic method is presented to standardize or codify addresses that need a large number of hand-coded data [14]. Nooj is a new corpus processing system with large-coverage multilingual dictionaries and grammars [19]. Taşkın et al. use Nooj for institution standardization [23]. The system sCooL is based on mappings collected from different curated and non-curated sources [9], for which knowledge from CareerBuilder (CB) and Wikipedia and manual mappings are critical. Google Scholar Citations (GSC) provides an institutional affiliation link by institution name and institutional e-mail web domain [16].

Pre-prepare rules and knowledge are valid for a given data set, which can be fully explored by statistic methods. However, when faced with large-scale data, the rules can be very complicated, and the knowledge needs to be extremely rich. For not pre-defined data sets (the number is increasing, and the type and distribution for institutions are changing with time), traditional approaches are undoubtedly limited by pre-prepare rules and knowledge.

All in all, the main limitation of the traditional methods lies in they are (1) difficult to scale up to increasing and changing affiliations; (2) hard to find new institutions, and (3) can’t explore the relationship between institutions. However, the development of knowledge graph technology provides us with a new idea to solve this problem. Our work focuses on solving the IND problem by entity linking with the help of existing knowledge graph.

III. PRELIMINARIES

Before introducing the framework, we present the preliminaries to systematically define the problems to be solved by our framework.

A. Xlore & XLink

X Lore\(^1\) is a large-scale multi-lingual knowledge graph by structuring and integrating Chinese Wikipedia, English Wikipedia, French Wikipedia, and Baidu Baike, which contains 16,284,901 instances, 2,466,956 concepts and 446,236 properties [24]. XLink is a large-scale English-Chinese bilingual knowledge graph based on X Lore, which provides external knowledge that can help readers to understand ambiguous and obscure entities [26]. The API of X Lore\(^2\) allow us to access the data expediently.

The most import concepts in knowledge graph are entity, attribute, and relation. For X Lore, all of them are prepared and easy to use. In this work, we mainly use instances, classes, and entities in X Lore. The outline of an Instances and a Classes shown in Figure 1 and Figure 2.

For each instance and class, there is a corresponding uri, which gives the details of the instance and the class (as it is shown in Figure 3).

An API query will return a set of instances and classes, which is based on fuzzy matching in X Lore. The right part of Figure 5 sketches their relationships (See API for their definitions and structures).

B. PROBLEM DEFINITION

First of all, we define the mapping relationship between external query w and internal entities as h, where \(h(w) = I + P, I_w\) is a collection of Instances and \(P_w\) is a collection of Classes.

---

1. https://xlore.org/
2. https://xloareapi.docs.apiary.io/api-description-document
And then, the goal of the entity linking based IND problem can be described as “find the exact top-level institution name from a set of entities”.  

Let $a \in \mathcal{A}$ as an affiliation string extracted from a scholarly paper, $o \in \mathcal{O}$ means an unambiguous top institution entity in the real world, $\mathcal{C}$ denotes the candidate set of $a$, where $\mathcal{C} \neq \emptyset$. According to our survey, the main points of the work includes:  
1) make sure the result entity is in the candidate set $\mathcal{C}$;  
2) selection algorithm needs to be accurate.  

The ultimate goal of the entity linking based method is to establish the mapping relationship between $\mathcal{A}$ and $\mathcal{O}$, which can be described as follows:

$$o = \arg \max_{c \in \mathcal{C}} P(c|a)$$  

In this work, the framework addresses two significant parts of the IND problem:  
1) **Candidate Generation**: generation of reliable candidate sets $\mathcal{C}$, taking all the possible IND factors mentioned above into consideration to ensure coverage of the candidate set to the result;  
2) **Result Selection**: a model to select the most likely result $o$ from $\mathcal{C}$, which will ensure accuracy of the linking between affiliation and the real-world top-level institution.

In general, the goal of our framework is to find the corresponding entities of affiliations by applying knowledge graph. Meanwhile, the framework can cope with the ever-changing and ever-increasing paper data, which implies all kinds of institutions.

**IV. FRAMEWORK**

To deal with above problems, we illustrate an institution name disambiguation framework based on entity linking in this paper. In this section, we specify the technical details of our framework.

**A. FRAMEWORK OVERVIEW**

As it is illustrated in Figure 4, the framework is consist of four parts: Preprocessing, Candidate Generation, Result Selection, and Application.

Generally, preprocessing of affiliations in papers can be roughly divided into three steps: paper structuring, affiliation extraction, and affiliation normalization. The principal aim of these steps is to group the same expressions into one to reduce processing times. What is more, the framework provides various of applications. In the next section, we will display two case studies based on our framework.

In the process, we make the most of the similarity between external strings and entities in the knowledge graph as well as the relationship between entities. To sum up, the framework mainly solves two problems: **Candidate Generation** and **Result Selection**, which are the most critical parts of the framework.

**B. CANDIDATE GENERATION**

Previous studies shows that a considerable number of errors in IND come from lacking of essential information in affiliation string. To overcome this difficulty, we use knowledge graph to learn its related expressions and add the related entities into $\mathcal{C}$ for selection.

The following aspects are mainly considered in the entity linking based **Candidate Generation**:  
1) institutions in $\mathcal{C}$ must be top-level and verifiable;  
2) all possible results must be included, including those not in the knowledge graph;  
3) any $c$ in $\mathcal{C}$ must be standard and accurate.
According to the above aspects, we propose a candidate generation algorithm based on entity linking supported by XLore. Before introducing the algorithm, we propose two crucial functions in this paper:

1. affiliation cleaning function \( f_1 \): The primary purpose of \( f_1 \) is removing the parts that do not contain any expression of an institution, which is similar to the traditional rule-based methods. We use \( s = f_1(a) \) to express the process. Meanwhile, we define the inverse process of \( f_1 \) as \( f_1' \), which will find the original affiliation string \( a \) of \( s \).

2. query string generation function \( f_2 \): \( f_2 \) will generate a contiguous sequence of \( n \) items from a given \( w \) by \( n \)-gram language model. The process is illustrated in the left part of Figure 5.

In \( n \)-gram language model, an \( n \)-gram is a contiguous sequence of \( n \) items from a given sequence of text or speech. In more detail, the number of elements in \( W \) (as the left part of Figure 5 shows, \( s \) contains \( m \) words) can be expressed as follows:

\[
|W| = \begin{cases} 
1, & m < n \\
\sum_{i=N}^{m} (m-i+1) = \frac{(m-N+1)(m-N+2)}{2}, & m > N
\end{cases}
\]  

(2)

Figure 5 depicts the process of linking an affiliation string \( a \) to a knowledge graph \( G \) in XLore. It's worth saying that an entity caching module is used to cache the query results from the knowledge graph, which can reduce the number of knowledge graph requests effectively.

where \( N \) denotes the lower limit of \( n \) in \( n \)-gram language model. Learning the latent expression in affiliation from knowledge graph not only gives us more background knowledge but also provide us with more relationships between entities and classes to better understand \( a \) semantically. In this way, we generate a reliable candidates \( C \) by the candidate generation algorithm.

The generation process of candidates set \( C \) of \( a \) is described as follows with pseudo-code:

\[
P(c|a) = P(c|s), \quad s = f_1(a) \\
P(c|a) = P(c'|s'), \quad c' = clean(c) \quad \text{and} \quad s' = clean(s)
\]

Obviously, this assumption is quite reasonable.
Algorithm 1 Candidate Generation Algorithm

Input: \( a \): affiliation string of a paper;  
Output: \( C \): candidate set of institutions;  

1. Initialize \( C \leftarrow \{\} \);  
2. \( s = f_1(a) \);  
3. \( W = f_2(s) \);  
4. for \( w \) in \( W \) do  
   5. add \( w \) into \( C \);  
   6. \( G = h(w) = I + P \);  
   7. for instance in \( I \) do  
      8. top-level institution identification, add label of instance into \( C \);  
   end  
8. for \( p \) in \( P \) do  
   9. top-level institution identification, add label of \( p \) into \( C \);  
   10. for instance in \( Instances \) of \( p \) do  
    11. top-level institution identification, add label of instance;  
   end  
12. for \( h \) in \( Hyponymys \) of \( p \) do  
   13. top-level institution identification;  
   end  
end

Based on this assumption and a series of experiments, we proposed a probability formula of entity linking, which can take full advantage of entity in knowledge graph and affiliation string. The probability formula can be expressed as follows:

\[
P_c = 1 - \frac{med(\text{cls}(c', s'), c' \cap s')}{\text{cls}(c', s')}  \quad (3)
\]

where \( \text{cls}() \) means a function to find the Longest Common Subsequence (LCS) [5] of two strings, \( \text{med}() \) denotes a function to calculate the Minimum Edit Distance (MED) [15] of two strings, \( C' \) suggests the mapping result of \( C \) by \( \text{clean}() \) function, and the absolute value represents the length of the string. It is worth mentioning that LCS is a NP-hard problem and we use dynamic programming approach to solve it [12].

In this paper, we use \( P_{c'} \) to describe the probability that \( s' \) and \( c' \) have the same meaning. Thus, we can get the following formula:

\[
\text{arg max } P(c'|s') = \text{arg max } P_{c'}, \quad c' \in C'
\quad (4)
\]

Furthermore, the result of formula 3 can be defined as follows:

\[
\mathcal{R}' = \text{arg max } P(c'|s')
\quad (5)
\]

where \( \mathcal{R}' \) denotes the result set of the entity linking. There is no doubt that \( \mathcal{R}' \) probably has more than one element. However, we simply choose the longest one \( r' \) and map \( r' \) to \( r \) with \( \text{clean}() \) function. In this way, we will get most possible candidate from \( \text{arg max } P(c'|a') \). The model denotes that the most possible candidate \( c \) given \( a \) depends on the text similarity between the top institutions.

The advantages of this method include: (1) combine the advantages of traditional methods; (2) make full use of the entities in the knowledge graph; (3) low computational complexity, which can deal with large-scale data.

V. EXPERIMENT

In this experiment, we mainly verify three main aspects: (1) the framework can achieve a considerable high accuracy; (2) the results have more knowledge than traditional methods; (3) it can deal with the ever-changing and ever-increasing data.

We designed a real-world experiment to verify the reliability and usability of our framework. To restore a more realistic application scenario, we randomly select papers from AMiner\(^3\) whose period is relatively long. The details of the data set are as follows:

For comparison, we also try to choose a state-of-the-art method. Actually, for rule-based methods and knowledge-based methods, the number of rules and the amount of knowledge determine the accuracy. Considering the completeness and generality of knowledge and rules, we deem that the work of Huang et al. [8] is most representative of all the papers mentioned above. Thus, we replicate their method on our data set and compare it with ELAD. In this paper, we call it “Huang’s Method” for short.

To be more specific, the data involved in ELAD are illustrated as follows:

In calculating Number per affiliation, we use the number of regularized affiliations. As Table 2 shows, our framework provides a large number of candidates for us to select, which provides the possibility of solving problems such as spelling mistakes, missing words, etc. In terms of the completeness of candidates, ELAD provides a higher possibility to improve accuracy.

| Items                       | Total Number | Number per affiliation |
|-----------------------------|--------------|------------------------|
| Papers                      | 100,000      | 5.11                   |
| Affiliations                | 201,038      | 5.02                   |
| Regularized affiliations    | 84,974       | 36.06                  |
| Labeled data                | 1,000        | 36.55                  |

\(^3\)https://www.aminer.org/
Finally, we designed a rigorous test method to verify the accuracy of our approach. In more detail, we only label the official name of an institution in this experiment, and only the exactly matched string is considered right. In addition, the usability of the framework enables us run the work at a very low cost.

VI. RESULTS
We apply these data in both ELAD and Huang’s method. Huang’s method clustered data into 10,714 categories, each of which represents an institution. With an average of 3.14 affiliations cluster into one institution, 33,628 of these affiliations are in clusters, which counts 39.57% of the whole. That’s to say, 40.23% affiliations did not involve in, which is much lower on our real-world data set than on their data set.

A. ACCURACY
From the perspective of accuracy, ELAD demonstrates clear superiority in this experiment. As Table 3 shows, the accuracy and recall of ELAD are much higher than the traditional method.

|               | Accuracy | Recall  | Return Result |
|---------------|----------|---------|--------------|
| Huang’s Method| 63.83%   | 31.65%  | 67.62%       |
| ELAD          | 74.80%   | 98.9%   | 100%         |

where Return Result means the ratio of results returned by the method.

B. MORE KNOWLEDGE
What’s more, with entity linking, ELAD can deal with university and university system level IND problem. For example, our framework deems the top institution of “Medical Anthropology Program, University of California, San Francisco, U.S.A.” is “University of California, San Francisco”, not “University of California”. Furthermore, some affiliations with a small amount of misspelling can also get the right result.

However, affiliations with severe lack of essential and correct information can’t be predicted, e.g., “Mochrurm, Newton Stewart, Wigtownshire, Scotland”. It is also to find the official name of “Inorganic Chemistry 2, Chemical Center, University of Lund, Lund, Sweden” is “Lund University”. Nonetheless, our framework has significantly improved recall and precision.

C. FOR EVER-CHANGING AND EVER-INCREASING DATA
During the preparation of experiment data, we find that the number of entities and uris did not increase linearly as the amount of data grow, which growth rate is decreasing. This phenomenon demonstrates that there are many common attributes among institutions, and ELAD is suitable for ever-changing and ever-increasing data processing.

Besides, the source code of ELAD, data set and experiment results are available at GitHub.4

VII. CASE STUDY
To validate our proposed novel framework, two cases studies are defined. In the first case study, we try to find out the most productive institutions in AI 2000 Data Mining Most Influential Scholars.5 We link the affiliation strings to their knowledge graph entities by applying ELAD. The top 10 institutions of the most productive are demonstrated in Table 4.

| Organization | Paper Number |
|--------------|--------------|
| Microsoft    | 1115         |
| University of Illinois Urbana Champaign | 1087         |
| IBM          | 1019         |
| Tsinghua University | 868         |
| Carnegie Mellon University | 794         |
| Chinese University of Hong Kong | 656         |
| Arizona State University | 623         |
| Yahoo!       | 558          |
| University of Illinois Chicago | 500         |
| Stanford University | 309         |

To the best of our knowledge, the result is consistent with the common-sense understanding, which verifies the reliability of ELAD in turn.

In the second case study, we try to mine the collaboration relationship between the top 10 institutions in the first case study. The collaboration frequency between institutions implies their relationships in academic cooperation and competition. The institutions at the centre of the collaboration network are often the most influential in this field. In this way, we visualize the collaboration network of the top 10 institutions in Figure 6.

4https://github.com/ShawJoe/elad.git
5https://www.aminer.cn/ai2000/dm
The nodes denote institutions, the edges imply the collaboration, and the values indicate the collaboration frequencies between institutions. The network proves the close cooperation between top research institutions.

**VIII. CONCLUSION**

This paper proposes an automatic framework to solve the IND problem in research papers in a scenario where the amount of data is increasing and the period is very long. Our framework makes full use of the relationships between entities in affiliations and their hypernym and hyponym, which can overcome the limitations of pre-preparation rules and knowledge methods. Experiments show that its recall is much higher than traditional methods with their precision at the same level. At the same time, our framework reduces human involvement with more external knowledge, which helps establish the linking between entities and give standardized names.

The framework makes it convenient for us to build a knowledge graph of research institutions, which includes the relationship between top institutions as well as the relationship between top-level institutions and sub-level institutions. Nevertheless, the most significant limitation is the speed of querying the knowledge graph data, which makes the framework more suitable for incremental processing. Meanwhile, the entity caching module in our framework can significantly reduce the frequency of queries.

**REFERENCES**

[1] G. Abramo, T. Cicero, and C. A. D’Angelo, “A field-standardized application of DEA to national-scale research assessment of universities,” *J. Informet.*, vol. 5, no. 4, pp. 618–628, Oct. 2011.

[2] D. Aumüller, “Towards Web supported identification of top affiliations from scholarly papers,” *Proc. Datenbanksys. Bus.*, 2009, pp. 1–12.

[3] D. Aumüller and E. Rahm, “Web-based affiliation matching,” in *Proc. ICIIQ*, 2009, pp. 246–256.

[4] K. C. Chan, A. Fung, H.-G. Fung, and J. Yau, “Ranking of institutions and academic journals,” *Managerial Finance*, vol. 42, no. 4, pp. 292–302, Apr. 2016.

[5] V. Chvátal and D. Sankoff, “Longest common subsequences of two random sequences,” *J. Appl. Probab.*, vol. 12, no. 02, pp. 306–315, Jun. 1975.

[6] P. Cuxac, J.-C. Lamirel, and V. Bonvallot, “Efficient supervised and semi-supervised approaches for affiliations disambiguation,” *Scientometrics*, vol. 97, no. 1, pp. 47–58, Oct. 2013.

[7] R. D. Bruin and H. Moed, *The Unification of Addresses in Scientific Publications*. Amsterdam, The Netherlands: Elsevier, 1990.

[8] S. Huang, B. Yang, S. Yan, and R. Rousseau, “Institution name disambiguation for research assessment,” *Scientometrics*, vol. 99, no. 3, pp. 823–838, Jun. 2016.

[9] F. Jacob, F. Iaved, M. Zhao, and M. McNair, “Scold.: A system for academic institution name normalization,” in *Proc. Int. Conf. Collaboration Technol. Syst. (CTS)*, May 2014, pp. 86–93.

[10] Y. Jiang, H.-T. Zheng, X. Wang, B. Lu, and K. Wu, “Affiliation disambiguation for constructing semantic digital libraries,” *J. Amer. Soc. Inf. Sci. Technol.*, vol. 62, no. 6, pp. 1029–1041, Jun. 2011.

[11] S. Jonnalagadda and P. Toplam, “Nemo: Extraction and normalization of organization names from PubMed affiliation strings,” *J. Biomed. Discovery Collaboration*, vol. 5, p. 50, Aug. 2010.

[12] S. N. Majumdar and S. Nechaev, “Exact asymptotic results for the Bernoulli matching model of sequence alignment,” *Phys. Rev. E, Stat. Phys. Plasma Fluids Relat. Interdiscip. Top.*, vol. 72, no. 2, Aug. 2005, Art. no. 020901.

[13] H. F. Moed and G. Halevi, “A bibliometric approach to tracking international scientific migration,” *Scientometrics*, vol. 101, no. 3, pp. 1987–2001, Dec. 2014.

[14] F. Morillo, J. Aparicio, B. González-Albo, and L. Moreno, “Towards the automation of address identification,” *Scientometrics*, vol. 94, no. 1, pp. 207–224, Jan. 2013.

[15] G. Navarro, “A guided tour to approximate string matching,” *ACM Comput. Surveys (CSUR)*, vol. 33, no. 1, pp. 31–88, Mar. 2001.

[16] E. Ou-Maulea, J. M. Ayllón, A. Martín-Martín, and E. D. López-Cózar, “The lost academic home: Institutional affiliation links in Google scholar citations,” *Online Inf. Rev.*, vol. 41, no. 6, pp. 762–781, 2017.

[17] F. Sabah, S.-U. Hassan, A. Muazzam, S. Iqbal, S. H. Soroya, and R. Sarwar, “Scientific collaboration networks in pakistan and their impact on institutional research performance,” *Library Hi Tech*, vol. 37, no. 1, pp. 19–29, Mar. 2019.

[18] Z. Shao, J. Tang, Y. Zhang, B. Gao, and Y. Wang, “Scholar trajectory: Visualizing the migration of talents,” Tsinghua Univ., Beijing, China, Tech. Rep. 201811149733.6, 2018.

[19] M. Silberzei, “Noof: A linguistic annotation system for corpus processing,” in *Proc. HTL/EMNLP Interact. Demonstrations*, 2005, pp. 10–11.

[20] H. Sun, J. Li, Y. Wu, L. Wang, and K. W. Fung, “Using an ontology-based approach to handle author affiliations in a large biomedical citation database,” *Stud. health Technol. Inform.*, vol. 245, p. 1338, 2017.

[21] H. Sun, L. Wang, Y. Wu, W. Hua, and J. Li, “Matching strategies for institution names in literature database,” *Data Anal. Knowl. Discovery*, vol. 2, no. 8, pp. 88–97, 2018.

[22] Z. Taşkin and U. Al, “Institutional name confusion on citation indexes: The example of the names of turkish hospitals,” *Procedia Soc. Behav. Sci.*, vol. 73, pp. 54–55, Feb. 2013.

[23] Z. Taşkin and U. Al, “Standardization problem of author affiliations in citation indexes,” *Scientometrics*, vol. 98, no. 1, pp. 347–368, Jan. 2014.

[24] Z. Wang, J. Li, Z. Wang, S. Li, M. Li, D. Zhang, Y. Shi, Y. Liu, P. Zhang, and J. Tang, “XLore: A large-scale English-Chinese bilingual knowledge graph,” in *Proc. Int. Semantic Web Conf. Posters Demos*, vol. 1035, 2013, pp. 121–124.

[25] H. Xu, “A regional university-industry cooperation research based on patent data analysis,” *Asian Social Sci.*, vol. 6, no. 11, p. 88, 2010.

[26] J. Zhang, Y. Cao, L. Hou, J. Li, and H.-T. Zheng, “Xlink: An unsupervised bilingual entity linking system,” in *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*. Beijing, China: Tsinghua Univ., 2017, pp. 172–183.

[27] Y. Zhang, F. Zhang, P. Yao, and J. Tang, “Name disambiguation in AMiner: Clustering, maintenance, and human in the Loop,” in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 1002–1011.

[28] H. Zhu, S. Jiang, H. Chen, and M. C. Rocco, “International perspective on nanotechnology papers, patents, and NSF awards (2000–2016),” *J. Nanoparticle Res.*, vol. 19, no. 11, p. 370, Nov. 2017.

ZHOU SHAO was born in Yichang, Hubei, China. He received the joint B.S. degrees from the Southwest University of Science and Technology and the University of Electronic Science and Technology of China, in 2011, and the M.S. degree in computer application technology from the Southwest University of Science and Technology, in 2014. He is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Nanjing University of Science and Technology. From 2014 to 2016, he was a Teaching Assistant with the Southwest University of Science and Technology. His research interests include data mining, knowledge graph, data analytics, and deep learning.
Xiangying Cao was born in Nantong, Jiangsu, China. She received the B.S. degree from Nanjing Forestry University, in 2019. She is currently pursuing the M.S. degree with the Information School, Renmin University of China. Her research interests include data mining and knowledge graph.

Sha Yuan received the B.S. degree from the Huazhong University of Science and Technology and the Ph.D. degree from the University of Chinese Academy of Sciences. From 2017 to 2019, she was a Postdoctoral Fellow with Tsinghua University. She is currently a Research Associate with the Beijing Artificial Intelligence Academy. Her research interests include data mining, deep learning, social networks, and the Internet of Things.

Yongli Wang (Member, IEEE) was born in Jiamusi, Heilongjiang, China, in 1974. He received the B.S. and M.S. degrees from the Harbin Institute of Technology and the Ph.D. degree in computer application technology from Southeast University, in 2006. He is Full Professor with the Nanjing University of Science and Technology. His research interests include database technology, knowledge graph, data mining, the Internet of things, and data analytics. He is a member of CCF and ACM.