An Evaluation Model for Authors' Academic Influence Based on Multi-source Heterogeneous Database in Bilingual Environment

Zhijian Deng, Zhenhua Zhang*, Chun Xiong, Mingqin Chen, Yanting Huang, Junran Li and Jiawen Shen

Guangdong University of Foreign Studies, School of Mathematics and Statistics, 510006, Higher Education Mega Center, Panyu, Guangzhou, China
Email: Zhenhua Zhang, zhangzhenhua@gdufs.edu.cn

Abstract. As the internationalization of academic rise up, it is insufficient to evaluate a scholar’s influence only taking domestic academic contribution as reference. To have a better understanding of scholars’ academic contribution in an international scale, this paper proposes a new evaluation system of academic influence based on the optimal dataset of entity recognition in bilingual environment. The new evaluation system has been experimented on a real literature dataset of DBLP. The research is conducted by the following steps. First, handle bilingual data. Second, optimize the literature data with entity recognition technology. Third, make an academic influence evaluation. Experiment results show that the new evaluation system is feasible and reasonable.

1. Introduction
In the era of big data, the literature data has grown into mass data causing a redundancy problem. Take the database of scientific and technological literature as an example, there is a high repetition rate of entities’ name, which brings great difficulties to distinguish the author's identity when searching on the database. Besides, there is still not commonly data integration for bilingual or multilingual literature database. Due to the differences between languages, entities matched wrong often occur when we try to make a bilingual data fusion. Worse, languages difference sometimes causes data silos. However, few studies of evaluation system of academic influence try to integrate bilingual or multilingual data, and the information to evaluate the scholar is incomplete, making the evaluation defective. A completed evaluation system of academic influence breaking through the obstacle of bilingual or multilingual data can give a fair and objective reference when quantify scholar’s academic influence, and hence it is urgently needed to be established. According to the existing research, a complete academic impact evaluation framework for scholars covers scholar entity recognition and academic impact assessment models.

The following is a review of the existing researches of entity recognition technology in the field of database. Wang and Fan (2011, [1]) summarized the research status of entity recognition technology on XML data, graph data and complex networks data. Zhao and Zhang (2014, [2]) studied the concepts and applications of several entity recognition technologies on XML documents. At present, the experimental data of entity identification technology is mainly derived from DBLP English document information and artificial data sets. However, there are few experiments on Chinese data sets. Schulz et al. (2014, [3]) used citation networks to deal with large-scale author name disambiguation. Korean clinical texts are used by Lee et al (2018, [4]) to compare...
the clinical named entity recognition (NER) method in Korean clinical literature of rheumatism patients, which verified that dictionary-based string matching and conditional random field (CRF) are better methods for implementing clinical NER in Korean clinical narrative documents. In recent years, with the spring up of big data, many scholars have begun to pay attention to the entity recognition technology on the big data platform. With a huge amount of data, the entity recognition technology based on Map-Reduce framework and Hadoop platform improves the performance of entity recognition. Huo and Wang (2013, [5]) proposed a big data entity recognition algorithm based on Map-Reduce framework, which calculates the similarity among records by attribute values, and then carried out entity recognition by graph clustering and experimented on Hadoop platform to prove the parallelism of the algorithm and the effectiveness and efficiency of processing big data. Qi et al (2016, [6]) proposed a Hadoop-based entity recognition algorithm of power big data attributes. The experiments prove the correctness of the algorithm and its advantages in discrete breakpoint numbers and acceleration ratio that is suitable for power data attributes. In the field of Chinese author identification, Xu (2019, [7]) studied the author identification of Chinese texts, and Wu (2019, [8]) studied the named entity recognition of multilingual migration.

There are a few related researches on the quality of database data in libraries, mainly focusing on topics, literature review, information collection, total quality management and other aspects of quality control and optimization of libraries, while the research literature specifically for the quality of bilingual libraries in both Chinese and English is even rarer. He and Liu (2011, [9]) explored the quality control of the whole process of database construction from six aspects: characteristic database planning, metadata, data source, recording standard control, retrieval function and quality control after database completion. Among them, the control of metadata redundancy is needed in the quality control of metadata. Wang (2014, [10]) proposed to introduce the theory of total quality management into library management, and control the quality of the characteristic database in university libraries from stages of planning, implementation, inspection and processing, so as to ensure the quality of characteristic database construction.

In the field of academic influence evaluation, Some scholars have carried out related research and have obtained some achievements. Wang (2019, [11]) revealed the distribution characteristics of the influence of scholar group papers under different scientific research evaluation systems through multi-dimensional index analysis. Cheng and Li (2020, [12]) divided the quantitative evaluation indicators into 4 types according to the citation frequency, citation and citation quality, citation time, and citation dispersion of academic journal papers, and proposed the evaluation indexes of future academic journals in view of the existing deficiencies in the four classifications. The above research is only focused on the evaluation of journals and papers. In the field of academic influence of scholars, Shi (2019, [13]) studied the evolution of influence of outstanding scholars based on academic big data and Liu (2019, [14]) studied the academic influence evaluation of scholars based on cooperation and citation network. However, at present, there is no authorship influence evaluation model in bilingual or multilingual environment. The existing evaluation indicators, such as the H index, the i10 index, and the total citation times of the authors' papers submitted by Google Scholar, both quantify the citations of the author's published papers. And there are some scholars choose social networks to analyze the journal bibliographic information. It is believed that researchers should cooperate in scientific research so as to improve the quality of research and promote the development of specific disciplines (2010, [15]). Nowadays, the more authoritative evaluation index of scholars' influence is the RG index of ResearchGate, a scientist's social networking website. RG Score, which takes into account the feedback and comments from other users of the Research Gate platform, is an effective index for evaluating the research performance of researchers. The existing research has laid a good foundation for the research of this paper. However, the above-mentioned indexes and models have not yet achieved the integration of Chinese and English. In real life, some scholars usually publish literature in Chinese and foreign journals. For these scholars, the existing evaluation system still is lack of comprehensiveness, accuracy and impartiality.

In summary, the current entity recognition technology has developed considerably, but there are few applications on the Chinese-English bilingual library database, and the existing academic
evaluation system of scholars urgently needs to perfect the deficiency information due to language differences.

By studying on the multi-source heterogeneous literature database, this paper focuses on the bilingual entity recognition technology and proposes a new evaluation system of academic influence. The simulation results show that this paper provides a fair and comprehensive model helping scholars have a better know for particular academic researches and exploit those valuable researchers.

This paper attempts to do some work in the evaluation of academic influence in a bilingual environment. The rest of the paper is organized as follows: The second part is the introduction to evaluation model of academic influence of bilingual literature database. The third part shows the specific process of bilingual database fusion and entity recognition. The fourth part presents a novel academic influence evaluation system for scholars and introduces the simulation experiment environment, the experiment data and the analysis of results. The fifth part is the conclusion and prospect of this paper.

2. Evaluation Model of Academic Influence of Bilingual Literature Database

![Flow chart of combining algorithm.](image)

In the process of integrating and fusing massive data, for the reason that there is no prescribed standard and unified expression, lots of data often have the difficulty of associating names with entities. The difficulty of association is manifested in the conflicts between naming and entity, which can be
divided into two categories, one is homonym, that is, one name corresponds to multiple entities; the other is synonym, that is, one entity has multiple naming references [9] (Li, 2015). Taking the literature author as an example, when searching for ‘Wang Wei’ in the literature engine such as HowNet, there will be many works by Wang Wei, but it is difficult for us to distinguish the different authors who are all called Wang Wei, which leads to the phenomenon of homonym. The synonym phenomenon refers to the fact that an author has two or more names, which may be named Wang Wei in Chinese, and his English names are ‘Wei Wang’ ‘W. Wang’ ‘Wang Wei’ or ‘Wang W.’ etc. The linguistic expressions of names in articles published by China and other countries will be inconsistent. At this time, entity recognition technology, also known as name disambiguation technology, is needed when we cannot establish the association between naming and entity. In this paper, we will study the bilingual entity recognition technology applied to document database. Considering the lack of research on bilingual database fusion in existing literature, The algorithm framework of this article mainly covers three aspects: bilingual data fusion, entity recognition and author influence evaluation. The specific content is shown in Figure 1 above.

2.1. Bilingual Database Fusion
At present, there are few studies on Bilingual database fusion, and there is no final conclusion on the methods of bilingual database fusion. Based on the existing research, this paper proposes the following two methods:

(1) Translate multi-lingual databases into the same language and then do the fusion, such as translating English data into Chinese and then do the fusion, or translating Chinese data into English and then do the fusion.

(2) The bilingual database is connected and fused by constructing parallel corpus.

In this paper, we first translate the names of all authors into Chinese and English to establish the mapping and correspondence between Chinese literature data and English literature data. And use keywords to establish a correspondence between all Chinese literature keywords and English translations. Based on the above operations, a bilingual parallel corpus with Chinese and English is created.

2.2. An Entity Recognition Integration Model Based on Text Clustering and Community Division
In terms of the characteristics of data, this paper uses text clustering algorithms and the community partitioning algorithm to construct an integration model.

(1) Text Clustering Algorithms.

In this paper, cosine similarity is used to calculate the similarity between objects, then the centroid method is used to calculate the distance between the classes (the arithmetic mean of all document vectors is used as the vector of the class). Finally, hierarchical clustering is used to cluster the data.

(2) Community Partition of Network Relations (Based on Graph Model).

In the identification of authors, authors tend to coauthor the authors who have coauthored in the past or closely related with. The characteristics of community structure are as follows: the density in the community is higher than that between communities, the inner connection of community is relatively close, and the connection between different communities is relatively sparse. Based on the above characteristics, network relations can be used for community partition of entity recognition. For community partitioning algorithm, modular index $Q$ is the current general standard, and its algorithm is to calculate the difference between continuous variables and expected values in each community. If the number of links in the community is higher than expected, it indicates that the nodes tend to concentrate in the community, that is, the modular structure of the network is more obvious. The formulas for calculating modular index $Q$ are as follows:

$$Q = \sum_{i=1}^{n}(e_{ii} - (\sum_{j=1}^{n}e_{ij})^2) = \sum_{i=1}^{n}(e_{ii} - a_{ij}^2)$$  (1)
\( n \) is the number of all lines (relationships) in the network, \( e_{ij} \) is the number of midlines in the first \( i \) community divided by the value of \( n \). \( e_{ij} \) is the sum of degrees of all nodes in the first \( i \) community and the number of relations of all nodes divided by the value of \( 2n \).

2.3. Evaluation Model of Academic Influence of Scholars

The research is divided into two parts: the personal academic evaluation model and the contributions in research team. The academic influence of scholars depends on the quantity, quality and contribution of published papers. The quality of the literature is measured by the journal impact factor of the year of publication. According to the signature order of the co-authors, the scholar's contribution to literature is calculated individually. The specific model will be introduced in the fourth part.

3. Process of Bilingual Database Fusion and Entity Recognition

3.1. The Source of the Data

In order to verify the effectiveness of framework we used, we select a large amount of document information and record data from the Chinese DBLP website (datatang.com). The data set is obtained from a project name “Innovation Method Group of Automation Subject”, which is provided by task group of the Institute of Automation in Chinese Academy of Sciences. The source of the English database is the DBLP public data set provided by the English DBLP website. Some of the scholar-related information and paper metadata of the academic papers published in the DBLP English website are used to constitute the experimental data set too. This paper focuses on the comprehensive evaluation of scholars in the field of computer science. In our study, in terms of the influencing factors of the journals, we selected 77 representative core journals in the computer field including Chinese and English journals. Among the identified scholar entities, scholars who published papers in the past five years (2014-2018) were selected as a sample, with a total of 428 scholars. In order to verify the entity identification framework, we select the 108 common names with the high frequency in the data set to form the training data set, which can be found in the appendix for data, and the other 320 names to form the testing data set.

3.2. Entity Recognition Process

The steps of the entity recognition process are as follows:

**Step 1**: Data cleaning.

In order to form the metadata for further experiment of we merge the meta data of articles under the same author's name together, which includes the fields title, keywords, abstract. Then text preprocessing is performed on the Chinese paper metadata, including translating English data into Chinese and text segmentation, and finally generating the original corpus of author meta data.

**Step 2**: Feature extraction.

After generating the corpus, a document-word word frequency vector matrix will be created. Then, through the TF-IDF weight calculation formula, we assign the TF-IDF weight of each word in the matrix, and then use the obtained TF-IDF value to sort and filter to further reduce the dimension of the vector space, so as to facilitate One-step text clustering.

**Step 3**: Text Clustering.

After obtaining the TF-IDF weight matrix, we choose to calculate the cosine similarity in order to perform systematic clustering Based on the assumption that each author sends two papers on average, the clustering results are divided into \( n/2 \) categories. At the same time, we will combine category and category by using the centroid method. Finally, the result of systematic clustering is re_so1.

**Step 4**: Extract the coauthor relationship.

For extracting author information that has a co-author relationship with the same author's name, we select the papers that the 108 authors selected above have participated in. Then by using the paperid as the unique identifier, we can from the table named “scholar-paper”, also known as the paperid-relationship table. From that table, we can extract the scholarid of all the authors who have co-authored the name.
Step 5: Community division. We draw the network diagram of the co-author relationship between the authors in the form of dotted lines, and the different nodes can be classified by the greedy algorithm, and the community to which they belong can also be quickly found. Finally the community division result of all the nodes can be obtained, denoted as re_so2.

Step 6: Combine the steps 4 and step 5 to generate the final result. Based on the systematical clustering method and community division, for the id set belonging to the same systematical group and belonging to the same community, we refer to this set as an entity. The result is recorded as Entity.

3.3. Simulation Experiment and Result Analysis
Based on the above steps, the training data sets are used to create an entity recognition model. And the results of text clustering and community division of eight author names with the highest frequency are shown in Table 1, we can evaluate the effect of entity recognition by its accuracy and recall rate. From the table below, we can tell that the accuracy of this experiment is high, which can tell from the fact that the average accuracy rate can reach 98.24%. All in all, the presented entity identification framework has high practical value for scholars' name disambiguation.

We have carried out a comprehensive entity classification on the basis of the above entity recognition algorithm, and the overall accuracy of the classification exceeds 90%. Next, we will carry out the establishment of the academic impact evaluation model of scholars based on the aforementioned entity recognition results.

| The name of author | Numbers of entity | Numbers of article | Accuracy | Recall rate |
|--------------------|-------------------|--------------------|----------|-------------|
| 王伟 Wei Wang       | 66                | 71                 | 96.58%   | 93.75%      |
| 张伟 Wei Zhang      | 58                | 60                 | 98.20%   | 92.21%      |
| 王勇 Yong Wang      | 54                | 57                 | 100%     | 95.32%      |
| 张军 Jun Zhang      | 46                | 48                 | 98.32%   | 91.03%      |
| 高峰 Feng Gao       | 39                | 43                 | 97.24%   | 87.24%      |
| 刘军 Jun Liu        | 41                | 43                 | 95.32%   | 90.23%      |
| 王平 Ping Wang      | 35                | 37                 | 97.63%   | 94.32%      |
| 王超 Chao Wang      | 30                | 35                 | 96.32%   | 95.32%      |

4. Scholarly Academic Influence Evaluation System
In order to evaluate the academic influence of the identified scholars, the influence factor will be used to measure the importance of the published papers. Besides, combined with the number of paper published by scholars and the contribution of scholars to paper, we can obtain an evaluation index based on literature measurement. Scholars’ academic influence is not only embodied in the literature, but also in the academic cooperation of scholars. Therefore, our study will combine social network and
evaluation indicators to weight the academic influence of scholars. Finally, the academic influence results obtained in our paper will be compared and analyzed with the RG Score of international mainstream ranking, so as to prove the rationality of the evaluation system in our paper.

4.1. Data Sources
The number of papers published by scholars, paper quality, social networking relationship and other information are evaluated comprehensively by scholars in computer science. In our study, 77 representative core journals in the computer field were selected, which based on the impact factor of journals. In the identified scholar entities, the scholars who published papers in the past five years (2014-2018 years) were selected as sample, a total of 428 scholars.

4.2. Academic Influence Evaluation Model
(1) Evaluation model of centrality degree based on social network.

Social network is mainly to study the relationship between social entities and the mode, structure and function of these connections. Generally, in order to achieve high academic attainments, scholars tend to cooperate academically with those who have high academic achievements. In scholar's co-author network, if a scholar has more cooperative relations with other scholars, then the scholar will own a greater centrality value, and hence it can be considered that the scholar is at the center of this social network and has greater academic influence. The degree of use of the undirected graph in scholar's co-author network to quantify the influence can be summarized as: The higher the degree of cooperation of scholars, the higher their core position, and the greater their academic influence.

The generated co-authors are connected with their cooperative relationship to form a co-occurrence matrix of co-authors, and are imported into the network analysis software UCINET according to the co-occurrence matrix to generate a network diagram of the author's cooperative relationship.

In the light of the co-occurrence matrix of co-authors, this paper uses $D_i$, the contact degree of scholar $i$, to express the centrality degree of scholar $i$. In fact, $D_i$ denotes the number of co-authors of
scholar $i$ in all his/her publications. In Figure 2, $D_i$ represents the number of edges in the $i$-th author's social network for co-authors. Thus $D_i$ can also be shown as follows:

$$D_i = \sum_{j=1}^{q} C_{ij}$$

(2)

Where $Q_i$ is the number of papers published by scholar $i$, $C_{ij}$ is the number of coauthors of scholar $i$ in paper $j$. When we calculate $C_{ij}$, we must delete the co-authors that have been counted among all the co-authors of the scholar $i$ in the previous paper from paper 1 to paper $j-1$. Calculate the centrality degree $D_i$ of author $i$ with the number of co-authors with different entities.

(2) Academic influence model of personal publications.

The academic influence of a scholar depends on the quantity, quality and contribution of the published papers. The quality of literature is measured by the journal impact factor $K_{ij}$ in the year of publication. The contribution of scholars to the literature is calculated harmoniously according to the signing order in the co-authored literature. The calculation formula is as follows:

$$W_{ij} = \frac{(1-r) \times r^{n_i-1}}{1-r^{N_i}}$$

(3)

Where $W_{ij}$ represents the contribution of scholar $i$ to coauthored paper $j$, $n_i$ denotes the order of authorship of scholar $i$ in the coauthored paper $j$, and $N_i$ is the total number of scholar $i$ who coauthored the literature. $r$ is a control variable, in order to make the value of the first author's rights in more than 30%, where the value is 0.7. According to the summary, $Q_i$ is the number of papers published by scholar $i$, $W_{ij}$ is the contribution of scholar $i$ to paper $j$, $K_{ij}$ is the impact factor of published journal, and the academic influence of the $i$-th scholar $F_i$ is calculated as follows:

$$F_i = \sum_{j=1}^{q} K_{ij} \times W_{ij}$$

(4)

(3) Integrated evaluation model.

The Min - Max method is used to standardize $F_i$ and $D_i$ to obtain $X_i$ and $Y_i$, and the final scholar $i$'s influence score was $G_i$ (results keep two decimal places).

$$G_i = \left( \frac{X_i + Y_i}{2} \right) \times 100$$

(5)

4.3. Analysis of Evaluation Results

On the basis of author entity recognition, our study explores the evaluation system of scholars' academic influence, which is based on the number of papers published by scholars, the quality of papers and the social network of scholars' cooperative relations. In addition, 428 scholars were ranked. Table 2 shows the top 10 scholars' influence scores in our study and partial scholars’ RG scores. As can be seen from the table, the author's academic influence ranking is roughly in line with RG Score ranking. The result is expected, which is to a degree to show the rationality of the evaluation system explored here.
Table 2. The top 10 authors of 428 scholars.

| ID  | Author          | F    | D    | X    | Y    | G    | RG Score |
|-----|-----------------|------|------|------|------|------|----------|
| 284 | Nirwan Ansari   | 13.20| 51.00| 1.00 | 1.00 | 100.00| 44.62    |
| 213 | Dianguo Xu     | 12.53| 50.00| 0.95 | 0.98 | 96.48 | 38.93    |
| 198 | Wei Wang       | 11.68| 44.00| 0.88 | 0.86 | 87.37 | 37.1     |
| 28  | Nengcheng Chen | 9.47 | 41.00| 0.72 | 0.80 | 76.08 | 34.61    |
| 226 | Feng Gao       | 7.37 | 49.00| 0.56 | 0.96 | 75.97 | 35.03    |
| 186 | Yijie Wang     | 11.67| 22.00| 0.88 | 0.43 | 65.77 | 32.92    |
| 183 | Yueshi Guan    | 12.42| 19.00| 0.94 | 0.37 | 65.67 | Missing  |
| 12  | Junyu Dong     | 5.45 | 44.00| 0.41 | 0.86 | 63.80 | 32.79    |
| 286 | Jun Liu        | 9.70 | 19.00| 0.73 | 0.37 | 55.37 | 30.99    |
| 374 | Heng Wang      | 8.25 | 17.00| 0.62 | 0.33 | 47.90 | Missing  |

5. Conclusion
After summarized current entity recognition technologies, considering the trait of the literature data, this paper makes an entity recognition which combines hierarchical clustering algorithm and community partitioning algorithm of the network relation. It’s proved that the presented entity recognition algorithm in this paper is accurate and valid. Based on the optimal dataset of entity recognition in bilingual environment, this paper has a new try on evaluation system of academic influence. Experiment results show that the system is reasonable.

This paper mainly explores a new evaluation system of academic influence in the bilingual environment, and finally we get a good result. However, there are still challenges in this field, and the system proposed in this paper is simple and can be improved: such as the diverse parallel corpus can be applied to the bilingual data fusion, realistic literature data set can be trained, accurate models to quantify a scholar’s contribution, etc.

6. Acknowledgments
This paper is funded by the National Statistical Research Key Project (No.2016LZ18), Guangdong Basic and Applied Basic Research Foundations (No.2018A030313470, 2016A030313688), Student Innovation Projects (No. S201911846035, S201911846029, 201511846058) of Guangdong Province, Teacher-Student Joint Research Project (No.18SS08) & Higher Education Project (No.2019GJ13Y) of Guangdong University of Foreign Studies.

7. References
[1] Wang H Z and Fan W F 2011 Object identification on complex data: A survey Chinese Journal of Computers vol 34 p 1843-1852
[2] ZHAO Z and ZHANG L C 2014 Research on entity identification technology on XML documents Computer Technology and Development vol 24(10) p 84-87.
[3] Schulz C, Mazloumian A, Petersen A M, et al. 2014 Exploiting citation networks for large-scale author name disambiguation EPJ Data Sci. vol 3(11) p 1-14
[4] Lee W, Kim K, Lee E Y and Choi J 2018 Conditional Random Fields for Clinical Named Entity Recognition: A Comparative Study Using Korean Clinical Texts Computers in Biology and Medicine vol 101 p 7-14
[5] Huo R, Wang H Z, Zhu R, Li J Z and Gao H 2013 Map-Reduce based entity identification in big data Journal of Computer Research and Development vol 50 p 170-179
[6] QI J, QU Z Y, LOU J L and WANG C 2016 A kind of attribute entity recognition algorithm based on Hadoop for power big data Power System Protection and Control vol 44(24) p 52-57
[7] Xu X L 2019 *Research on author identification of Chinese texts* (Beijing: People’s Public Security University of China)

[8] Wu H Q 2019 *Research on unsupervised named entity recognition based on cross-language migration* (Harbin: Harbin Institute of Technology)

[9] He J X and Liu F 2011 Discussion on quality control of library characteristic database *New Century Library* vol 5 p 68-71

[10] Wang L 2014 Quality control research on characteristic database of university library from the perspective of total quality management *Library Work and Study* vol 7 p 105-108.

[11] Wang K 2019 Research on the influence of journal articles of domestic and foreign crop scholars *Library and Information Studies* vol 12 (4) p 94-98, 87

[12] Cheng J J and Li X X 2020 Academic Journal Impact Evaluation Index and Improvement Scheme *Information Studies: Theory & Application* vol 3 p 1-10

[13] Shi Y J 2019 *Research on the evolution of influence of outstanding scholars based on academic big data* (Dalian: Dalian University of Technology)

[14] Liu K Y 2019 *Research on academic influence evaluation of scholars based on cooperation and citation network* (Jinan: Shandong Normal University)

[15] WEI S P 2015 The semantic model of AID and its realizing methods *Modern Educational Technology* vol 33(6) p 106-109