Advances in Economics, Business and Management Research (AEBMR), volume 54
2018 5th International Conference on Management Science and Management Innovation (MSMI 2018)

Research on the Classification Rules of Database Indexes in Author Name Disambiguation

Yue ZHOU[1,a], Tian LI[1,b] , Lu-Yao LIU[1,c], Yu-Ying ZHONG[1,d], Zheng-Lu YU[2,e], and Jun-Peng YUAN[3,4,f,*]

[1]Central University of Finance and Economics, Beijing, 102206, China

[2]Institute of Scientific and Technical Information of China, Beijing, 100038, China

[3]National Science Library, Chinese Academy of Sciences, Beijing,100190, China

[4]School of Economics and Management, University of Chinese Academy of Sciences, China

[a]2015311588@email.cufe.edu.cn, [b]2015311569@email.cufe.edu.cn, [c]2016311603@email.cufe.edu.cn, [d]2015311575@email.cufe.edu.cn, [e]luluyu@istic.ac.cn, [f]yuanjp@mail.las.ac.cn

*Corresponding author

Keywords: Author name disambiguation, Indicator analysis, Priority framework.

Abstract. In recent years, with the involvement of more scholars in academic research and rapid development of digital libraries, author name disambiguation has been one of the most critical problems. In this paper, we study how to make the best use of the author's personal information to distinguish different authors from a mixture of records with the same author's name. We construct an indicator evaluation system and a priority framework and establish a series of appropriate classification rules. After data validation, our classifier not only fits the characteristics of author information data, but also maximizes the value of author's personal information to solve the problem of author name disambiguation.

Introduction

Author name disambiguation is an acknowledged problem to be solved urgently. The complexity of the problem stems from the rapid development of digital libraries and the increasing involvement of scholars in academic research. The issue of duplicate name has led to the decline of the quality of document management and the decrease of document retrieval speed.

In recent years, many international scholars have come up with solutions based on different perspectives. Bibliographic data, collaborator information and citation information are common foundations of mainstream setup algorithms. For example, according to the bibliographic data of Chinese documents, Y.X. Zhu put forward a two-step decomposition and one-step merger framework based on rules and similarities, focusing on the author's "agency" [1]. J. Lang started from the character social network, and design clustering algorithm based on web search results [2]. B. Wu proposed a complete set of duplicate name dispelling system (NDC) with three step to ensure a higher accuracy of name disambiguation [3]. Q. Lin summarized the rules of artificial disagreement and creatively designed the user feedback collection method to improve the perceptron with human wisdom [4]. W.Q. Song proposed a better stepwise clustering method [5]. In addition, His research also explored the feasibility of a feature-graph approach.

A.A. Ferreira proposed a three-step unsupervised self-training approach to the author name disambiguation-SAND (Self-Training Association Name Disambiguation) to address the lack of information [6]. W.L. Liu designed an author name disambiguation system composed of similarity estimation and clustering for the biomedical paper database PubMed [7]. Y.N. Qian and A.F. Santana paid more attention to how to deal with newly added duplicate author data [8,9]. J. Schulz studied the quality of datasets generated by different Author Name Disambiguation Processes and used different Monte Carlo simulations to analyze the general impact of different literatures [10].

In general, in order to solve the problem of author name disambiguation, the main algorithms

studied by scholars are heuristic, unsupervised and clustering algorithms. However, the simple clustering algorithm has two shortcomings in dealing with the problem of duplicate name disambiguation. Firstly, when the classical clustering algorithm is applied to the problem of duplicate name disambiguation, the selection and extraction of features will become very difficult. Secondly, the clustering center of the problem we studied is not obvious, and the clustering algorithm can hardly achieve better performance. Therefore, in this study, we transform the clustering problem of the name disambiguation problem into a classification problem to determine whether two authors with the same name are the same entities. J. Kleinberg and S. Basu has applied a similar approach, that is, to find a classifier [11,12]:

$$f(i,j)=\begin{cases} 1, \text{Record i, j belong the same author entity} \\ -1, \text{Record i, j belong to different author entity} \end{cases} \quad \text{i, j is the record number} \quad i \neq j \qquad (1)$$

It would not be difficult to match a true author entity if we got the proper classifier to cluster a large number of author papers with the same name by controlling a loop. Our research focuses on the construction of a good classifier and evaluating the indicators, constructing a priority framework and forming the classifier rules based on identification of the author's identity in order to achieve the purpose of identifying authorship.

Indicator Analysis

Data Source

The data used in this study were from CNKI (China National Knowledge Internet). We select some authors whose own name has serious name duplication to construct author information database. The original data information is presented in two-dimensional table. Each article record contains the authors' Name, Gender, Year of Birth and many other text and numerical information, a total of 40 indicators.

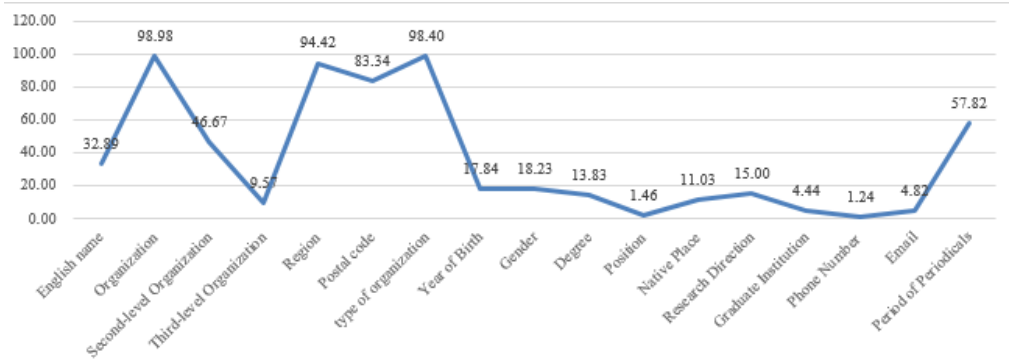

Fig. 1 Information integrity of the indicator [%]

Deterministic Indicators

According to the characteristics and effects of different indicators for judging whether the two records belong to the same author, we construct an index evaluation system to classify the indicators and consider the type of indicators that need to be empowered. The index evaluation system is as follows:

Deterministic indicators are divided into positive and negative deterministic indicators.

Positive deterministic indicators: these indicators show that only the indicator can be uniquely identified as the author of the same name, if the two records have same performance under these indicators, they must belong to the same author. If not, it does not help to distinguish authors. For example, two records have same Email, the corresponding must be the same author. Positive certainty refers to the contribution of the same performance of an indicator to the author disambiguation.

Negative deterministic indicators: characterized by the use of these indicators to distinguish different authors, if the two records have different performance under these indicators, you can surely deny that two records belong to the same entity. For example, two records of different origin, the

corresponding must be different author entity. Negative determinism refers to the contribution of different performance of an indicator to identify authors. It is noteworthy that the negative deterministic indicator do help when two records have same performance as well. Taking native place as an example, if the two records are of the same Native Place, the probability that the two records belong to the same entity are increased, which helps to determine the author's identity.

Uncertainty Indicators

Uncertainty indicators: No matter whether two records are the same or different under these indicators, it is not possible to judge whether two articles are from the same author or not.

Although a single negative certainty indicator or uncertainty indicator is not enough to completely determine the author's identity and effectively merge the different records, it still has a certain contribution to the identification of duplicate authors. At this time, we consider the selection of multiple indicators from the portfolio and use performance under the combination of indicators to determine the source of the two records. For example, after repeated tests, we get the indicator group of origin + year of birth + gender, which can correspond to the author of an essay with high accuracy.

For the indicator in the valid indicator combination, it is necessary to weight $\omega \in (0,1)$ respectively. (Note: It is not necessary to weight the positive certainty, and rules can be set to take advantage of the positive certainty.) If the two records perform the same under the category, the indicator scores a factor of 1; if not, the score is -1, and the total score is the sum of each indicator's score multiplied by the weight of the indicator. The two records have same performance under the k_{th} indicator.

$$S(i,j)=\sum_{k=m}^{n} \omega_k s_k,\ s_k \in \{1,-1\},\ i \neq j. \tag{2}$$

$$s_k=\begin{cases} 1, \text{ The two records have same performance under the } k_{th} \text{ indicator} \\ -1, \text{ The two records have different performance under the } k_{th} \text{ indicator} \end{cases} \tag{3}$$

Where *i* and *j* represent two records, k represents the indicator number from m to n (the first m-1 indexes are the positive certainty index), ω_k represents the weight of the indicator, and s_k represents the score of the two records under the indicator. The value range is $-1 \leq S(i,j) \leq 1$.

Classification of Indicator

Build an Indicator Set

A: author information indicator set, $A=\{a_k\}$

ω_k: a_k's weight

After a preliminary selection, some of the meaningless indicators and repeatability indicators were filtered out, and 17 indicators were retained from the original 40 indicators. Specific indicators in set A are as follows:

Table 1 Specific indicator in set A

a_1	English Name	a_{10}	Degree
a_2	Organization	a_{11}	Position
a_3	Second-level Organization	a_{12}	Native Place
a_4	Third-level Organization	a_{13}	Research Direction
a_5	Region	a_{14}	Graduate Institution
a_6	Post Code	a_{15}	Phone Number
a_7	Type of Organization	a_{16}	Email
a_8	Year of Birth	a_{17}	Period of Periodicals
a_9	Gender		

After preliminary indicator test with data, we consider the accuracy of information indicator, data type, degree of ambiguity, and the validity of the indicators. 8 indicators are ignored and will not be used as the main basis for the next set of combinations.

The updated index A set of indicators contains nine indicators, is divided into four categories:

1. Positive deterministic indicator set A_1 - Contact: a_1 (Phone Number), a_2 (Email)

2. Negative deterministic indicator set A_2 - Personal Status: a_3 (Year of Birth), a_4 (Gender), a_5 (Native Place)
3. A_3 - Work Status: a_6 (Organization), a_7 (Second-level Organization)
4. A_4 - History Information: a_8 (Degree), a_9 (Graduate Institution)

Priority Structure

Combining these four categories of indicators, giving full consideration to the integrity of each indicator, we set the priority for implementing the ratings.

According to the order of priority to determine whether the two records are from the same author. Processing rules are as Fig.2. The highest priority (priority 1) is the two indicators of the contact method, namely a_1 (Phone Number), a_2 (Email). The two indicators of the contact method are positive deterministic indicators, and without the help of other indicators, an author entity can be identified from the author information database. Sub-priority (priority 2) is three indicators of personal status, that is, a_3 (Year of Birth), a_4 (Gender), a_5 (Native Place). They are all negative deterministic indicators. For the same author entity, the three indicators will not change. The last priority (priority 3) is the work status and history information. Both types of indicators are indicators of uncertainty, but important for identifying authors with duplicate names.

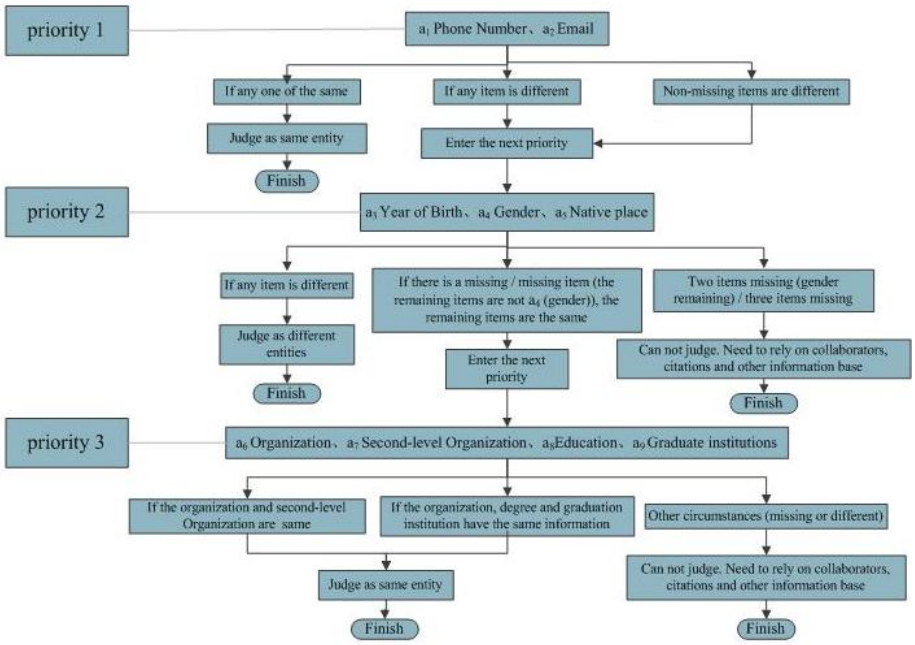

Fig. 2 Priority structure

According to the discriminant rules in the priority framework, we can get the weight ω of the corresponding indicator which constitutes the combination through algorithm training, calculate $S(i,j)$, judge the similarity of record, and then set the reasonable threshold x and y. On the basis of obtaining the weight ω, the algorithm can also explore other effective index combinations.

Summary

With the rapid development of digital library, more scholars participate in academic research, and the method of author name disambiguation becomes important. In this paper, according to the lack of data in various indicators of the database and data information, we analyze the role of indicators in the identification of duplicate names. Firstly, we select indicators with high utilization rate according to the completeness of the indicator content, then classify the indicators according to the indicator content, and finally divide the selected indicators into different levels. According to the results of the analysis, we can sort out a multi-level algorithm of author name disambiguation.

Acknowledgement

This research was financially supported by the National Natural Science Foundation of China (Grant NO. 71473236).

References

[1] Y.X. Zhu, Research on the Problem of Eliminating Author's Name Disambiguation of Chinese Bibliographic Data[J], Library and Information Service, 2014, 58(23):143-148+142.

[2] J. Lang, B. Qin, W. Song, L.Liu, T. Ting, S. Li, Name Search Result Disambiguation Based on Social Network[J], Journal of Computer, 2009, 32(07):1365-1374.

[3] B. Wu, C.L. Xu, W.B. Wang, W. Wu, Research and Application of Dealing With Author 's Duplicate name based on Link[J], Computer Science, 2008,(03):197-199.

[4] Q. Lin, Research on Algorithm of Ranking and Discrimination in Academic Networks[D], Huazhong University of Science and Technology, 2011 .

[5] W.Q. Song, Scientific and Technological Literature Authors Duplicate Name Disambiguation and Entity Link[D], Harbin Institute of Technology, 2012.

[6] A.A. Ferreira, A. Veloso, M.A. Gon çalves, et al, Self-training Author Name Disambiguation for Information Scarce Scenarios[J], Journal of the Association for Information Science & Technology, 2014, 65(6):1257–1278.

[7] W. Liu, D. R Islamaj., S. Kim, et al, Author Name Disambiguation for PubMed[J], Journal of the Association for Information Science & Technology, 2014, 65(4):765–781.

[8] Y.N. Qian, Q. Zheng, T. Sakai, et al, Dynamic Author Name Disambiguation for Growing Digital Libraries[J], Information Retrieval Journal, 2015, 18(5):379-412.

[9] A. F. Santana, A.H.F. Laender, A.A Ferreira., Incremental Author Name Disambiguation By Exploiting Domain-specific Heuristics[J]. Journal of the Association for Information Science & Technology, 2017, 68(4):931-945.

[10]J. Schulz, Using Monte Carlo Simulations To Assess the Tmpact of Author Name Disambiguation Quality on Different Bibliometric Analyses[J], Scientometrics, 2016, 107(3):1283-1298.

[11]J. Kleinberg, E. Tardos, Approximation Algorithms for Classification Problems with Pairwise Relationships: Metric Labeling and Markov Random Fields[C],Foundations of Computer Science, 1999. Symposium on. IEEE, 1999:14-23.

[12]S. Basu, A. Banerjee, R.J. Mooney, Active Semi-Supervision for Pairwise Constrained Clustering[J]. 2004:333--344.