Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 3790176, 7 pages
https://doi.org/10.1155/2021/3790176

Research Article

Chinese Personal Name Disambiguation Based on Clustering

Chao Fan [1,2] **and Yu Li** [1,2]

[1]*The School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, China*
[2]*Jiangsu Key Laboratory of Media Design and Software Technology, Jiangnan University, Wuxi 214122, China*

Correspondence should be addressed to Chao Fan; fanchao@jiangnan.edu.cn

Received 9 April 2021; Revised 26 April 2021; Accepted 29 April 2021; Published 15 May 2021

Academic Editor: Shan Zhong

Personal name disambiguation is a significant issue in natural language processing, which is the basis for many tasks in automatic information processing. This research explores the Chinese personal name disambiguation based on clustering technique. Preprocessing is applied to transform raw corpus into standardized format at the beginning. And then, Chinese word segmentation, part-of-speech tagging, and named entity recognition are accomplished by lexical analysis. Furthermore, we make an effort to extract features that can better disambiguate Chinese personal names. Some rules for identifying target personal names are created to improve the experimental effect. Additionally, many calculation methods of feature weights are implemented such as bool weight, absolute frequency weight, tf-idf weight, and entropy weight. As for clustering algorithm, an agglomerative hierarchical clustering is selected by comparison with other clustering methods. Finally, a labeling approach is employed to bring forward feature words that can represent each cluster. The experiment achieves a good result for five groups of Chinese personal names.

1. Introduction

The ambiguity of named entities is a prevalent phenomenon in natural language. There is considerable ambiguity about the personal name in the texts or the web pages, especially in the Chinese dataset. The Chinese personal name "Gao Jun (高军)" has a total of 51 items in the Baidu Encyclopedia. Eliminating the ambiguity of such personal name is beneficial to many tasks like information retrieval and data summarization. Take searching a person name on the Internet for example, documents of different person entities with the same name can be found by search engine. It is necessary to divide the documents into clusters automatically and secure the key information of each cluster. This research focuses on this task of importance and attempts to solve the problem by unsupervised approaches.

Chinese personal name disambiguation involves distinguishing between people with an ambiguous name in Chinese corpus. Initially, documents with the html format in raw corpus are processed into plain texts. Then, the lexical analysis of documents is performed, including segmentation, part-of-speech tagging (POS tagging), and named entity recognition (NER). Feature selection is enforced according to the result of lexical analysis. In order to acquire better accuracy of personal name recognition, some rules of personal name extension are proposed for target names to be disambiguated. For instance, the family name and first name of a target name may be separated in some situation due to the segmentation errors. We merge them into one complete personal name with the purpose of reducing the number of discarded documents. Further, an agglomerative hierarchical clustering algorithm is adopted to discover different clusters containing the same personal name. Finally, the label of each cluster is given by scoring the weight of each feature word in cluster. The feature words chosen as the cluster label can represent the person entity with significant information.

The rest of this article is arranged in the following parts. Related work of this task is introduced in Section 2. Research framework and methodology are elaborated in Section 3. Section 4 gives the experimental results and some discussions. Conclusion and future work are discussed in Section 5.

2. Related Work

The personal name disambiguation task is similar to the word sense disambiguation (WSD). Both of them pursue

the goal of resolving the ambiguity in natural language understanding. Nevertheless, there is a big difference between two tasks. The number of person entities with an ambiguous personal name is usually unknown for the name disambiguation task, which is contrary to WSD. Hence, personal name disambiguation is often implemented with an unsupervised clustering.

There are many research directions in personal name disambiguation. Song et al. [1] exploited two topic-based models to extract features from corpus and achieved a good effect for personal name disambiguation. Zhao et al. [2] made use of the personal ontology to complete feature extraction and similarity calculation on two real datasets, where the highest similarity is selected for disambiguation. Xu et al. [3] utilized a network embedding-based technique to disambiguate the author name, in which networks are created from papers that have a target ambiguous author name. Yu and Yang [4] solved the challenging task under the circumstances of inadequate data sources. A feature learning means and an affinity propagation clustering were taken into account. Kim et al. [5] combined global features with structure features for author name disambiguation. Global features, extracted from attributes of dataset, formed the textual vector representation. Moreover, negative samples were employed to train a global model. Protasiewicz and Dadas [6] produced a hybrid framework considering both rule-based method and agglomerative hierarchical clustering. Rules were generated from the knowledge of experts, analysis, and so forth. A function C_{index} was also proposed to determine the best threshold for stopping the hierarchical clustering algorithm. Du et al. [7] applied spectral clustering to recognize ambiguous names in large-scale scientific literature datasets. A distributed approach using Spark framework was advanced to perform in large-scale datasets. Pooja et al. [8] concentrated on the namesake issue of author name disambiguation. They presented an ATGEP method by taking advantage of a graph theory combined with an edge pruning operation.

The research on personal name disambiguation in Chinese datasets is also studied among a number of scientists. Chen et al. [9] provided a feature weighting scheme by calculating pointwise mutual information between personal name and feature word. A trade-off indicator was designed to measure the quality of clusters and stop hierarchical clustering. Li and Wang [10] developed a multistage clustering algorithm based on the entity knowledge, which can be used for Chinese named entity recognition and disambiguation. Ke et al. [11] handled the author name disambiguation under the condition of insufficient information and missing data. Their algorithm devised a novel combination of indicator and incorporated back propagation neural networks.

3. Framework and Methodology

3.1. Research Framework. The research framework of disambiguating personal name is depicted in Figure 1. There are four main parts in the framework: preprocessing, feature selection, clustering, and labeling. The detailed procedure can be presented in the following steps.

(1) Preprocess the raw corpus by removing the html tags

(2) Perform the lexical analysis, including Chinese word segmentation, POS tagging, and named entity recognition

(3) Select words as features and build feature vectors to represent documents

(4) Chose the weighting scheme for feature vectors

(5) Calculate similarity between documents and perform the clustering algorithm

(6) Evaluate the clustering results and assign each cluster a label

3.2. Feature Selection. Extracted features should be capable of distinguishing between people with the same name. The features selected in this paper are outlined as follows:

(i) Feature 1: named entities (NE)

(ii) Feature 2: nouns (N)

(iii) Feature 3: nouns and verbs (N + V)

(iv) Feature 4: nouns with their document frequency $(\text{df}) > 1$ (N + Df1)

(v) Feature 5: named entities with name extension (NE + NameEx)

(vi) Feature 6: nouns with name extension (N + NameEx)

(vii) Feature 7: nouns and verbs with name extension (N + V + NameEx)

(viii) Feature 8: nouns with their df > 1 and name extension (N + Df1 + NameEx)

where df represents the count of documents having a certain term. A word with df = 1 should be ignored since it cannot contribute to the discrimination between documents.

Due to the weakness of the Chinese word segmentation tool, the personal name in the document may not be correctly identified. A series of rules for name extension are devised according to the results of word segmentation, which is shown in Table 1. Part-of-speech "nr," "nr1," "nr2," "nrf," and "ng" represent "personal name," "Chinese surname," "Chinese given name," "transcribed personal name," and "noun morpheme." "w" denotes "punctuation mark." The extension of personal name improves the accuracy of target name recognition.

Also, feature selection can be performed in the whole document (Document) or the paragraphs (Paragraph) encompassing the target personal name. Different schemes give birth to different results.

3.3. Feature Weight. As documents are represented by a vector space model (VSM), each feature vector can be obtained by calculating the feature weight of a document. A variety of weighting schemes are raised in previous work. This

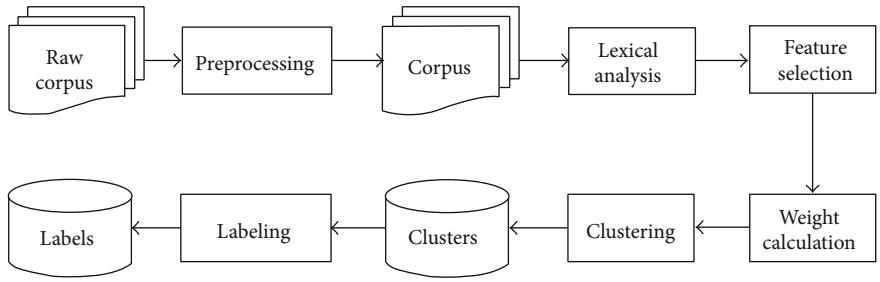

Figure 1: Research framework of personal name disambiguation.

Table 1: Rules for personal name extension (nr is a Chinese part-of-speech representing personal name).

Rule no.	Pattern	Examples
Rule 1	nr1 + n = > nr	Li(李)/nr1 Jun(军)/n => Li Jun(李军)/nr
Rule 2	nr + nr2/ng = > nr	Li Jun(李军)/nr Shi(师)/ng => Li Junshi(李军师)/nr
Rule 3	nrf + w + nrf = > nr	Roger(罗杰)/nrf./w Musson(穆森)/nrf => Roger Musson(罗杰.穆森)/nr

research adopts four different types of weight calculation methods, which are discussed in detail as follows.

(i) *Boolean Weights*. A weight assigned to a word is either 0 or 1.0 represents absence of a word in the document, whereas 1 represents the presence. Formula (1) displays the mathematical expression of Boolean weights, where f_{ij} indicates the frequency of word i existing in document j.

$$w_{ij} = \begin{cases} 1 & \text{if } f_{ij} > 0, \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

(i) *Frequency Weights*. This weighting scheme gives each word an absolute frequency, which is the number of occurrences of word i in document j.

$$w_{ij} = f_{ij} \quad (2)$$

(i) *tf-idf Weights*. A word that appears in only a few documents is likely to be a better discriminator than one that occurs in most or all documents. Inverse document frequency (idf) gives greater weight to words that appear in fewer documents. The tf-idf weight assigned to word i in document j can be calculated by formula (3).

$$w_{ij} = tf_{ij} \times \text{idf}_i = tf_{ij} \times \log \frac{N}{\text{df}_i}, \quad (3)$$

where tf_{ij} is f_{ij} divided by total number of words in the document. N is the count of documents in entire collection and df_i is the number of documents with word i

(i) *Entropy Weights*. The entropy weight method introduces the concept of entropy to measure the distribution of words i in document j, so the basic idea of the entropy method is similar to idf. It can be defined as follows formula (4):

$$w_{ij} = \log\left(tf_{ij} + 1\right) \times \left(1 + \frac{1}{logN} \sum_{j=1}^{N} \frac{tf_{ij}}{n_i} \log\left(\frac{tf_{ij}}{n_i}\right)\right) \quad (4)$$

3.4. Clustering Algorithm

3.4.1. Hierarchical Clustering. Hierarchical clustering can be divided into two types: divisive and agglomerative. This paper chose the latter because the complexity of divisive clustering algorithm is relatively high and not practical for this task.

Agglomerative hierarchical clustering belongs to a bottom-up method [12]. It treats each document containing target personal name as a separate cluster in the beginning. The algorithm merges two most similar clusters into a larger one at each step until the maximum similarity of clusters exceeds a preset threshold or there is only one cluster left. For the similarity formula, this paper calculates cosine of the angle between the vectors $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$ and $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)$ [13]. It can be written as formula (5).

$$sim(\boldsymbol{x}, \boldsymbol{y}) = \cos(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \bullet \boldsymbol{y}}{|\boldsymbol{x}| \bullet |\boldsymbol{y}|} = \frac{\sum_{i=1}^{n} x_i \bullet y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \bullet \sum_{i=1}^{n} y_i^2}}. \quad (5)$$

3.4.2. K-Means Clustering. K-means algorithm selects k documents as centroids to form initial clusters. Each document is

Table 2: Data statistics for each personal name in corpus.

Personal name	Number of documents	Discarded (gold standard)	Discarded (ICTCLAS)	Discarded (LTP)
Li Jun (李军)	234	1	0	4
Roger (罗杰)	357	24	3	2
Gao Jun (高军)	300	82	84	107
Sun Ming (孙明)	207	2	2	2
Zhang Jianjun (张建军)	247	0	0	1

repeatedly assigned to different clusters according to the closest centroid. Then, the centroid of each cluster will be recomputed. The iteration stops when a convergence criterion is satisfied or after a fixed number of iterations.

3.4.3. Spectral Clustering. Spectral clustering is a type of graph-based clustering. It utilizes the eigenvalues or spectrum of the similarity matrix to achieve the goal of dimensionality reduction. Documents can be assigned to different clusters based on the lower-dimensional representation. There are three basic stages in spectral clustering, including preprocessing, decomposition, and grouping.

3.4.4. GMM Clustering. Gaussian mixture models (GMM) clustering, also known as expectation-maximization (EM) clustering, makes use of the optimization strategy to cluster unlabeled documents. GMM assumes that data are generated by a Gaussian distribution and tries to obtain a mixture of multidimensional Gaussian probability distributions which can best model any dataset.

3.5. Labeling Approach. In order to summarize the person information of each cluster produced by the clustering algorithm, a labeling step is necessary. A simple way of creating a label is to choose a group of representative feature words by ranking the weights of all feature words in cluster.

The labeling algorithm [9] combines mutual information (MI) with *tf* to score the weights. For each feature word x_i in cluster C_k, the score is calculated by formula (6). $\mathrm{MI}(x_i, \mathrm{name})$ measures the mutual information between the feature word and personal name. $tf(x_i, C_k)$ counts the number of x_i appearing in cluster C_k. We can acquire a label of k words by taking the top k feature words in the scoring process.

$$\begin{aligned}\mathrm{score}(x_i, C_k) = {} & \mathrm{MI}(x_i, \mathrm{name}) \times \mathrm{MI}_{\mathrm{name}}(x_i, C_k) \\ & \times (1 + \log(tf(x_i, C_k))),\end{aligned} \tag{6}$$

$$\mathrm{MI}(x_i, \mathrm{name}) = \frac{p(x_i, \mathrm{name})}{p(x_i) \times p(\mathrm{name})} = \frac{\mathrm{df}(x_i, \mathrm{name}) \times N}{\mathrm{df}(x_i) \times \mathrm{df}(\mathrm{name})}, \tag{7}$$

$$\mathrm{MI}_{\mathrm{name}}(x_i, C_k) = \frac{p(x_i, C_k)}{p(x_i) \times p(C_k)} = \frac{\mathrm{df}(x_i, C_k) \times N}{\mathrm{df}(x_i) \times df(C_k)}. \tag{8}$$

4. Experiment and Discussion

4.1. Dataset. The dataset (the dataset is from CLP-2010.) is composed of 1345 files with html tags, including 109 discarded documents that do not have correct target personal names. The contents of documents are from "People's Daily." There are five personal names and each of them contains 200-400 news corpus.

Table 3: Comparison of clustering algorithms (feature 1 + paragraph + tf).

Clustering algorithm	Precision	Recall	F score
Hierarchical	78.04%	89.76%	**80.94%**
K-means	77.44%	77.68%	74.94%
Spectral	68.03%	91.74%	74.16%
GMM	77.34%	82.24%	77.91%

Chinese word segmentation, parts-of-speech tagging, and named entity recognition are performed on corpus. Two types of segmentation and tagging toolkits are exploited: ICTCLAS (http://ictclas.nlpir.org/) and LTP(http://ltp.ai). As the performance of word segmentation has an important impact on the accuracy of personal name recognition, we compared the number of discarded documents for two toolkits (see Table 2). Gold standard gives the real number of discarded documents. Result shows the personal name recognition of ICTCLAS is more precise than LTP.

4.2. Evaluation. Purity and inverse purity [14, 15] are taken as precision and recall for evaluating the clustering effect. Suppose S is the cluster set to be evaluated and R is the manually labeled category set. The definition of purity and inverse purity can be described by formula (9) and (10).

$$\mathrm{Precision} = \mathrm{Purity} = \frac{\sum_{S_i \in S} \max_{R_j \in R} |S_i \cap R_j|}{\sum_{S_i \in S} |S_i|}, \tag{9}$$

$$\mathrm{Recall} = \mathrm{InversePurity} = \frac{\sum_{R_i \in R} \max_{S_j \in S} |R_i \cap S_j|}{\sum_{R_i \in R} |R_i|}. \tag{10}$$

The F score calculates the harmonic mean of precision and recall, which is defined by formula (11). The overall F score of five personal names is the average of these values.

$$\mathrm{Fscore} = \frac{2\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}. \tag{11}$$

5. Result and Discussion

The Chinese personal name is recognized by the lexical analysis tool ICTCLAS in this research. We classify all documents containing the same personal name into one

Table 4: Comparison of feature selection (tf-idf).

Features	Document			Paragraph		
	Precision	Recall	F score	Precision	Recall	F score
NE	80.15%	92.24%	83.32%	77.74%	88.84%	80.36%
N	81.38%	93.74%	84.59%	82.49%	92.85%	85.15%
N + V	80.10%	93.83%	83.83%	80.93%	93.11%	83.86%
N + Df1	82.70%	92.76%	85.45%	82.77%	90.40%	84.75%
NE + NameEx	89.32%	92.55%	90.87%	86.81%	88.94%	87.79%
N + NameEx	91.07%	94.22%	**92.58%**	91.29%	93.12%	**92.15%**
N + V + NameEx	89.57%	94.31%	91.82%	90.55%	93.45%	91.92%
N + Df1 + NameEx	91.09%	93.25%	92.11%	90.28%	90.65%	90.40%

Table 5: Comparison of feature weight calculating approach (with highest F score).

Feature weights	Document			Paragraph		
	Precision	Recall	F score	Precision	Recall	F score
Bool (N + NameEx)	86.28%	93.99%	89.8%	90.66%	94.59%	**92.51%**
Tf (N + NameEx)	89.90%	93.44%	91.59%	91.00%	93.39%	92.11%
Tf-idf (N + NameEx)	91.07%	94.22%	**92.58%**	91.29%	93.12%	92.15%
Entropy (N + Df1 + NameEx)	90.54%	92.49%	91.47%	89.63%	86.88%	88.04%

directory and discard documents that do not contain a target personal name. This paper selects NE features involved in paragraphs containing the target name and frequency weights for testing different clustering approaches. The number of clusters k is extracted as a prior value from the gold standard. Results of four clustering algorithms are displayed in Table 3.

From Table 3, the hierarchical clustering algorithm outperforms other methods. Additional experiments utilizing other features showed similar results and verified the advantage of hierarchical clustering.

After choosing the clustering algorithm, different combinations of features are adopted for comparison. Table 4 summarizes the clustering results with tf-idf weights.

Named entities (NE) are reckoned as the baseline in this experiment. The F score increases significantly when other noun features are included (NE belongs to noun). However, adding verb features will lead to a drop in F score in comparison with only using noun features. Even though some verbs help to identify the identity of person, most verb features are very limited in disambiguating the personal name. A large number of unrelated verbs will bring noisy data in features, leading to poor experimental results.

Owing to the inaccuracy of word segmentation tools, personal names in documents cannot be effectively identified. When the recognition of an ambiguous name fails, documents that should be clustered are discarded. Therefore, personal name extension is introduced by setting some rules to identify target names. The results presented in Table 4 suggest that name extension dramatically improves the F score of clustering results.

The scope of feature selection can be either in the document or in each paragraph having the ambiguous name. According to Table 4, the feature selection in the whole document yields better results than in paragraph with the exception of noun feature.

Table 6: Result of personal name disambiguation (N + NameEx +document+tf-idf).

Personal name	Precision	Recall	F score
Li Jun (李军)	84.19%	86.32%	85.24%
Roger (罗杰)	82.20%	92.10%	86.87%
Gao Jun (高军)	100.0%	98.62%	99.31%
Sun Ming (孙明)	94.63%	98.52%	96.54%
Zhang Jianjun (张建军)	94.33%	95.55%	94.94%
Average	91.07%	94.22%	92.58%

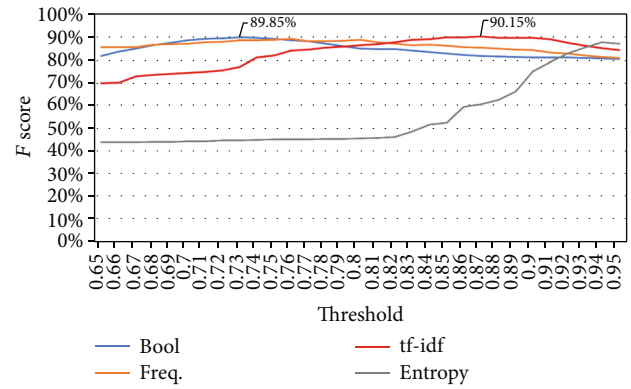

Figure 2: Distribution of F score at different thresholds.

Results of four feature weighting schemes are shown in Table 5 with highest F score. The corresponding features are listed in parenthesis in the first column. As can be seen

Table 7: Labels for "Gao Jun(高军)" clusters.

Entity	Number of documents	Produced labels
Gao Jun 1	208	选手(player), 金牌(gold medal), 冠军(champion), 奥运会(Olympic games), 朝鲜(North Korea), 邓亚萍(Deng Yaping), 队(team), 乔红(Qiao Hong), 女子(women), 乒乓球(table tennis)
Gao Jun 2	4	大队(brigade), 柬埔寨(Cambodia), 金边(Phnom Penh), 工程兵(engineer), 华人(Chinese), 桥梁(bridges), 大队长(captain), 运动会(games), 工兵(engineer), 李金勇(Li Jinyong)
Gao Jun 3	1	编辑(editor), 劳有林(Lao Youlin), 评论(comment), 李东生(Li Dongsheng), 吉菲(Ji Fei), 窦文涛(Dou Wentao), 民委(civil affairs committee), 大桥(bridge), 电视(TV), 杜晓春(Du Xiaochun)
Gao Jun 4	2	王圣珍(Wang Shengzhen), 选集(anthology), 深情(affection), 毛选(Mao Xuan), 字字句句(words and sentences), 写字台(writing desk), 手抄本(manuscripts), 春秋(spring and autumn), 日日夜夜(days and nights), 书法(calligraphy)

from the table, results obtained by tf weight in the whole document are better than others.

The detail result of every personal name is described in Table 6 when the average F score for all names achieves a best value.

Clustering algorithm stops when the similarity between two clusters is less than a certain threshold. The relationship between threshold and F score can be illustrated in Figure 2. The influence of threshold on results depends on feature sets. We choose feature 6 (N + NameEx) plus document to run clustering. The value of threshold is given through enumeration by every 0.01 step. The F score reaches a highest value of 90.15% when tf-idf weight is selected.

Basic information about a person is given by labeling process. For instance, clusters of Gao Jun(高军) are labeled with meaningful words in Table 7. The created labels are representative words that can summarize the characteristics of a person.

6. Conclusions

This paper studied the task of Chinese personal name disambiguation based on an unsupervised method. The open dataset contains five ambiguous names with gold standard. We exploited lexical analysis toolkits to perform segmentation and POS tagging. Eight groups of features are selected to combined with four feature weight calculating methods. In order to refine the precision of personal name recognition, name extension is proposed. The extension process of personal name significantly enhances the final effect of clustering experiments. Besides, the agglomerative hierarchical clustering algorithm is chosen from four methods for disambiguating names. The threshold of hierarchical clustering is also tested for different feature weights. At last, labels are constructed for clusters of target name by scoring the weights of feature words in clusters.

Final experimental results demonstrated the effectiveness of the proposed research approach. Nonetheless, some disadvantages may exist in the framework. Rules of personal name extension are suited for the current dataset. It may be necessary to add extra rules for other corpus so as to increase the precision of detecting Chinese personal names. In addition, we will develop automatic feature selection algorithms as well as new weigh calculating methods in future work. More sophisticated clustering and supervised document classification methods will also be taken into consideration.

Data Availability

The original dataset used in this work is available from the corresponding author on request.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work was supported by the Youth Foundation of Basic Science Research Program of Jiangnan University, 2019 (No. JUSRP11962) and High-level Innovation and Entrepreneurship Talents Introduction Program of Jiangsu Province of China, 2019.

References

[1] Y. Song, J. Huang, I. G. Councill, J. Li, and C. L. Giles, "Efficient topic-based unsupervised name disambiguation," in *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, pp. 342–351, Vancouver, BC, Canada, 2007.

[2] Z. Lu, Z. Yan, and L. He, "Ontology-based personal name disambiguation on the web," in *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, pp. 185–192, Atlanta, GA, USA, 2013.

[3] J. Xu, S. Shen, D. Li, and Y. Fu, "A network-embedding based method for author disambiguation," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 1735–1738, Torino, Italy, 2018.

[4] Z. Yu and B. Yang, "Researcher name disambiguation: feature learning and affinity propagation clustering," in *International Symposium on Methodologies for Intelligent Systems*, pp. 225–235, Springer, 2018.

[5] K. Kim, S. Rohatgi, and C. L. Giles, "Hybrid deep pairwise classification for author name disambiguation," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 2369–2372, Beijing, China, 2019.

[6] J. Protasiewicz and S. Dadas, "A hybrid knowledge-based framework for author name disambiguation," in *2016 IEEE*

International Conference on Systems, Man, and Cybernetics (SMC), pp. 000594–000600, Budapest, Hungary, 2016.

[7] H. Du, Z. Jiang, and J. Gao, "Who is who: name disambiguation in large-scale scientific literature," in *2019 International Conference on Data Mining Workshops (ICDMW)*, pp. 1037–1044, Beijing, China, 2019.

[8] K. M. Pooja, S. Mondal, and J. Chandra, "A graph combination with edge pruning-based approach for author name disambiguation," *Journal of the Association for Information Science and Technology*, vol. 71, no. 1, pp. 69–83, 2020.

[9] C. Chen, H. Junfeng, and W. Houfeng, "Clustering technique in multi-document personal name disambiguation," in *Proceedings of the ACL-IJCNLP 2009 Student Research Workshop*, pp. 88–95, Suntec, Singapore, 2009.

[10] G. Li and H. Wang, "Chinese named entity recognition and disambiguation based on multi-stage clustering," *Journal of Chinese Information Processing*, vol. 27, no. 5, pp. 29–34, 2013.

[11] H. Ke, T. Li, Y. Zhou, Y. Zhong, Z. Yu, and J. Yuan, "Aauthor name disambiguation using BP neural networks under missing data," *Journal of the China Society for Scientific and Technical Information*, vol. 37, no. 6, pp. 600–609, 2018.

[12] C. Schätzle and H. Booth, "DiaHClust: an iterative hierarchical clustering approach for identifying stages in language change," in *2019 in Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pp. 126–135, Florence, Italy, 2019.

[13] Q. Zhou and L. Leydesdorff, "The normalization of occurrence and co-occurrence matrices in bibliometrics using cosine similarities and Ochiai coefficients," *Journal of the Association for Information Science and Technology*, vol. 67, no. 11, pp. 2805–2814, 2016.

[14] C. Fan and J. Yu, "Finding community structure in social network of Renren," *ICIC Express Letters*, vol. 7, no. 5, pp. 1693–1698, 2013.

[15] A. Hotho, S. Staab, and G. Stumme, "WordNet improves text document clustering," in *Proceedings of the SIGIR 2003 Semantic Web Workshop*, pp. 541–544, Toronto, Canada, 2003.