Personal Name Disambiguation for Chinese Documents in Online Medium

Chao Fan[1,2(✉)] and Yu Li[1,2]

[1] The School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, China
fanchao@jiangnan.edu.cn

[2] Jiangsu Key Laboratory of Media Design and Software Technology, Jiangnan University, Wuxi 214122, China

Abstract. Disambiguating various people that share the same name is a critical issue for analyzing contents in online medium. This paper develops a framework for dealing with personal names in Chinese dataset. Web pages containing personal name are crawled from the online website and standardized at first. Then documents are parsed with lexical analysis technologies, such as segmentation, part-of-speech tagging, named entity recognition. We extract several groups of words as features, testing different weighting schemes (e.g. Boolean term frequency, absolute term frequency, tf-idf, entropy weights). By conducting the agglomerative clustering, a measure of interdependence within clusters and independence between clusters is proposed for automatically determining the number of clusters. Moreover, a technique that merges noise clusters is utilized to improve the clustering results. Experiments are performed on six groups of Chinese personal names and the final results confirm our proposed approach.

Keywords: Personal name disambiguation · Chinese personal names · Agglomerative clustering

1 Introduction

Personal name disambiguation plays a significant role in analyzing documents on the Internet. When searching a person's name online, we are confused with the problems that the search results often contain web pages of different people with the same name. Therefore, it is necessary to disambiguate names in the retrieved documents by cluster analysis. Furthermore, personal name disambiguation can be applied in the areas such as information collection, information fusion, etc. Exploring the key issue will increase the convenience for the end users and ultimately benefit the content and service providers.

This paper makes an effort to investigate the Chinese personal name disambiguation. Firstly, web pages are crawled from Baidu news via specified personal name keywords. By removing the web page tags, the text body is extracted with a standardized format. Some irrelevant documents are deleted and the rest are manually tagged with labels, which are used for scoring the algorithm. Secondly, we segment all documents with

W. Fu et al. (Eds.): ICMTEL 2021, LNICST 387, pp. 255–264, 2021.
https://doi.org/10.1007/978-3-030-82562-1_23

a segmentation tool, tagging the parts of speech and recognizing the named entities. Six groups of words extracted as features are combined with four weighting models, attempting to reach the best combination for feature selection. Thirdly, an agglomerative clustering algorithm is implemented employing selected features. Further, an indicator is adopted to measure the performance of clustering and the final number of clusters is optimized by removing the noise clusters. The dataset created in this paper is composed of documents with six different Chinese personal names. Each name contains 100 Chinese documents.

The contents can be organized in the following steps. Section 2 introduces related work of personal name disambiguation. Section 3 gives the framework, dataset and evaluation method of this work. Section 4 and 5 display the most important parts of feature generation, clustering approach and experiment results. In Sect. 6, conclusion and future work are discussed.

2 Related Work

The task of personal name disambiguation draws attention from a number of researchers. Li et al. [1] proposed a multi-stage clustering algorithm when the entity knowledge base is provided. Zhao et al. [2] constructed personal ontology and calculated the similarity between ontology and instance built from web pages. Emami [3] extracted semantic information from web pages and exploited a graph-based algorithm to disambiguate personal names. Xiong et al. [4] solved the personal name disambiguation problem based on the sentential semantic structure analysis. Yang et al. [5] presented an algorithm based on ensemble. The method integrates different divisions produced by different clustering algorithms and shows a high accuracy and robustness. Zeng et al. [6] adopted the multi-feature fusion method to disambiguate expert with the same name in the process of building expert database. Shang et al. [7] discussed a disambiguation method based on co-authors and their affiliates. Zhai et al. [8] employed sparse distributed representation to disambiguate English author name. The summary is chosen as disambiguation feature in their research. Yu et al. [9] combined different network embedding methods to eliminate the ambiguity of author names. Pooja et al. [10] identified ambiguous author names applying a graph combination with edge pruning-based approach. Kim et al. [11] introduced a hybrid deep pairwise method by exploiting both structure and global features.

Some variants of this task are also concerned among different scholars. Delgado et al. [12] studied this problem in a multilingual context with real search results. Their method performs better than the translation ways. Khabsa et al. [13] addressed two practical issues: adding constraints and allowing the data to be added partially. These constraints can improve the disambiguation result.

As it is typically completed by an unsupervised method, the previous works concentrate on many aspects such as feature selection, similarity calculation, clustering algorithm, etc. This paper pays attention to traditional personal name ambiguation issue on Chinese dataset and explores both feature selection and clustering algorithm.

3 Overview

3.1 Framework

The framework of Chinese personal name disambiguation proposed in this paper can be depicted in Fig. 1. There are three main modules: corpus building, feature generation and clustering module. Web pages are crawled and standardized into original dataset when building corpus. Features used for clustering can be extracted from corpus via feature generation module. Clustering module applies the hierarchical clustering algorithm to obtain results of cluster analysis.

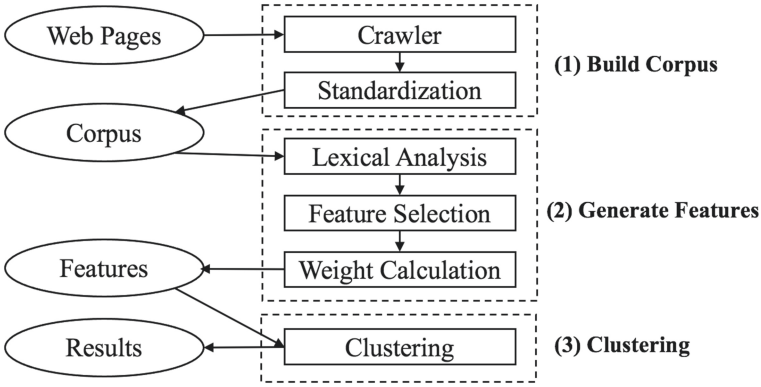

Fig. 1. Framework of Chinese personal name disambiguation

3.2 Dataset

The dataset is constructed in the corpus building module. We crawled Baidu news by keywords of ambiguous personal name. They are in Chinese characters and consist of the following six personal names: 陈卫 (CHEN Wei), 陈坚 (CHEN Jian), 刘伟 (LIU Wei), 刘伟强 (LIU Weiqiang), 李鹏 (LI Peng), 李小鹏 (LI Xiaopeng).

Web pages are parsed by removing the HTML tags at first. Standardization is then done by extracting the text body. Through this step, raw web pages are saved in a text-only format. We checked all documents manually and removed irrelevant ones. Eventually, 100 documents are kept for each personal name. In addition, documents in the dataset are tagged with labels as gold standard for the purpose of scoring the clustering algorithm.

3.3 Evaluation

P-IP score [14] is utilized to evaluate the clustering result of six personal names. The precision of a cluster P ∈ P for a given category L ∈ L is given by:

$$Precision(p,\ l) = \frac{|P \cap L|}{|P|} \tag{1}$$

According to the Precision formula, the Purity, Inverse Purity and F-score can be calculated for evaluation.

4 Feature Generation

4.1 Data Preprocessing

The documents in dataset are preprocessed with natural language processing toolkit LTP[1]. Stop words are filtered for texts of documents at the beginning. Then all sentences are segmented into words which are tagged with parts of speech. They are organized as such forms "江南(Jiangnan)/ns 大学(university)/n 校长(president)/n 陈卫(CHEN Wei)/nh …" where character and part of speech are separated with "/". Named entities can also be recognized by LTP with a high accuracy. Named entity often denotes real-world object like a person, an organization, a place and so forth. In this case, "江南大学(Jiangnan University)/Ni 陈卫(CHEN Wei)/Nh" will be identified as named entities. Jiangnan University (江南大学) is a name of organization and CHEN Wei (陈卫) is a Chinese personal name.

Words, parts of speech, and named entities are extracted from corpus at the first step of feature generation, which are prepared for feature selection and weight calculation.

4.2 Feature Selection

Different type of words or named entities can be selected as features. The simplest way is to incorporate all words as features of clustering. In this paper, six groups of features are chosen for testing the performance of hierarchical clustering, which are defined as follows.

Feature 1: all words except for punctuation;
Feature 2: all nouns;
Feature 3: all named entities;
Feature 4: all words with their document frequency (df) $>= 2$;
Feature 5: all nouns with their df $>= 2$;
Feature 6: all named entities with their df $>= 2$.

where document frequency (df) is the number of documents containing a particular term. Since the word with a df of 1 only appears in one document, it cannot distinguish between documents when running clustering. Thus, word with a df of 1 should be neglected even though it appears many times in only one document.

4.3 Feature Weight Calculation

It is necessary to choose the weighting scheme for features, because different words contribute differently to discriminating the documents. Four weight calculation approaches are provided in this paper, which can be elaborated on the following parts.

[1] http://ltp.ai.

Boolean Weights. It will be filled with 1 if a document has a word, otherwise it will be 0. It is calculated by formula (2), where f_{ij} is the frequency that word i appears in document j.

$$w_{ij} = \begin{cases} 1 & if\ f_{ij} > 0 \\ 0 & otherwise \end{cases} \tag{2}$$

Frequency Weights. The absolute term frequency f_{ij} counts the frequency of word i contained in document j.

$$w_{ij} = f_{ij} \tag{3}$$

Tf-idf Weights. Term frequency-inverse document frequency (tf-idf) considers both the number of a word in the document and in the corpus. It can be written as the following formula (4).

$$w_{ij} = tf_{ij} * log\frac{N}{n_i} \tag{4}$$

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

Entropy Weights. Entropy weights utilize the formula of entropy to reflect the distribution of words in the document, which can be defined as formula (5).

$$w_{ij} = log\left(tf_{ij} + 1\right) * \left(1 + \frac{1}{logN}\sum_{j=1}^{N}\frac{tf_{ij}}{n_i} log\left(\frac{tf_{ij}}{n_i}\right)\right) \tag{5}$$

4.4 Experiments of Feature Generation

The experiments combining six groups of features and four weighting schemes are performed for six Chinese personal names in the dataset. The best F-score value is selected from each experiment by testing all thresholds of the hierarchical clustering algorithm. Results of "陈卫 (CHEN Wei)" and "刘伟强 (LIU Weiqiang)" are shown in Figs. 2, 3, 4 and 5. The rest personal names exhibit similar characteristics.

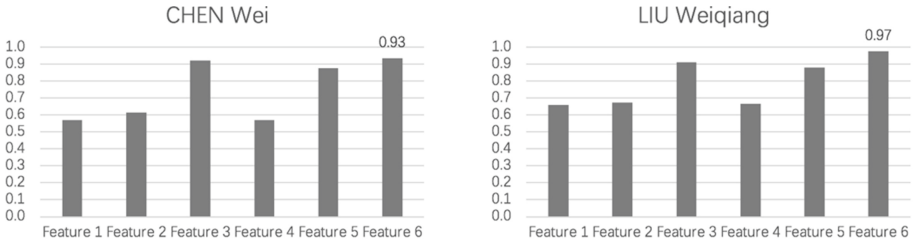

Fig. 2. Results of boolean weights for "CHEN Wei" and "LIU Weiqiang"

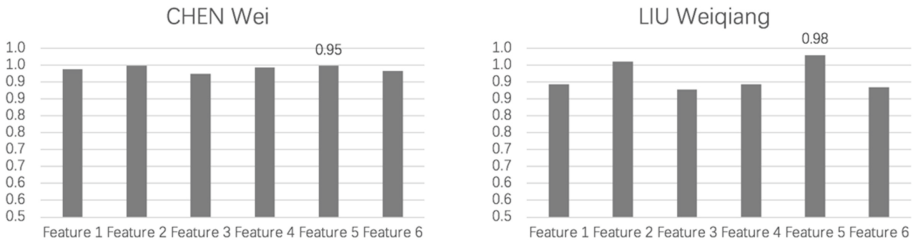

Fig. 3. Results of frequency weights for "CHEN Wei" and "LIU Weiqiang"

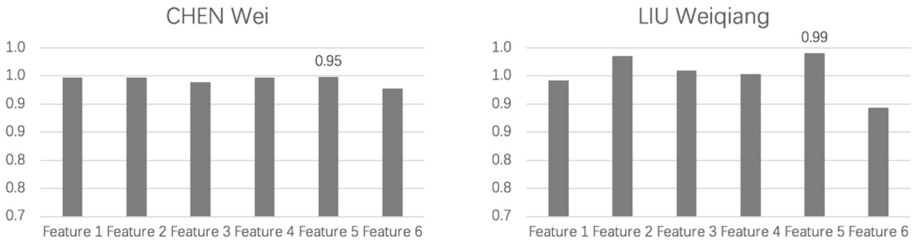

Fig. 4. Results of tf-idf weights for "CHEN Wei" and "LIU Weiqiang"

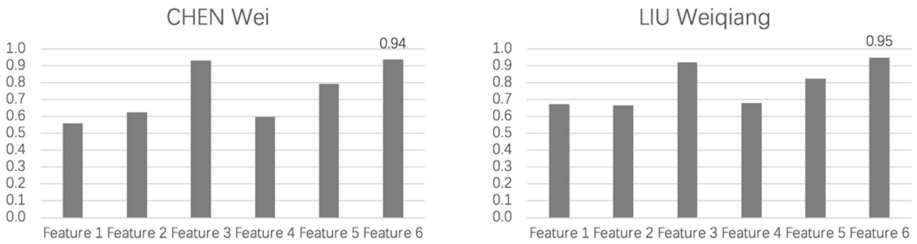

Fig. 5. Results of entropy weights for "CHEN Wei" and "LIU Weiqiang"

According to figures above, feature 6 has a relatively good effect on boolean and entropy weight. In contrast, feature 5 performs well on frequency and tf-idf weight. On average, feature 5 (all nouns with df $>= 2$) and tf-idf weights reach the best performance, so they are selected for cluster analysis.

5 Clustering

5.1 Hierarchical Clustering

An agglomerative hierarchical clustering algorithm is utilized in our personal name disambiguation task. In the first place, each document is reckoned as a single cluster. The similarity (or distance) for each pair of clusters is calculated at each step. Algorithm greedily selects the pair that achieve a greatest similarity and merge two clusters into a bigger one. In this work, the cosine similarity is adopted to measure the similarity

between two vectors $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$. It can be described as the formula (6).

$$sim(\boldsymbol{x}, \boldsymbol{y}) = cos(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{|\boldsymbol{x}| \cdot |\boldsymbol{y}|} = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \cdot \sum_{i=1}^{n} y_i^2}} \tag{6}$$

The clustering process is stopped when the similarity between two most similar clusters is smaller than a threshold. In our experiments, the average performance of six personal names achieves the best when the threshold of clustering is set to 0.81.

5.2 Cluster Stopping Measure

Newman devised a modularity Q indicator [15, 16] in community detection, which is used to measure the community structure and identify communities automatically. A high modularity Q denotes dense intra-community links and sparse inter-community links. According to this idea, the number of links can be replaced by the similarity of documents in a hierarchical clustering. Dense intra-community links indicate high intra-cluster similarity, whereas sparse inter-community links suggest low inter-cluster similarity.

Suppose that $C_{\boldsymbol{x}}$ and $C_{\boldsymbol{y}}$ are two of the m clusters obtained by clustering algorithm. $sim(\boldsymbol{x}, \boldsymbol{y})$ is the similarity between document $\boldsymbol{x}$ and document $\boldsymbol{y}$. The similarity between two clusters can be defined as follows.

$$e_{i,j} = \frac{\sum_{c_{\boldsymbol{x}} \in C_i} \sum_{c_{\boldsymbol{y}} \in C_j} sim(\boldsymbol{x}, \boldsymbol{y})}{\sum_{c_{\boldsymbol{x}} \in C} \sum_{c_{\boldsymbol{y}} \in C} sim(\boldsymbol{x}, \boldsymbol{y})} \quad i, j = 1, 2, \ldots, m \tag{7}$$

The intra-cluster similarity can be represented by $e_{i,i}$ $(i = 1, 2, \ldots, m)$. Hence, the clustering performance measure can be defined by a modularity Q as formula (8).

$$Q = \sum_{i=1}^{m} \left[e_{i,i} - \left(\sum_{j=1}^{m} e_{i,j} \right)^2 \right] \quad i, j = 1, 2, \ldots, m \tag{8}$$

As it measures the quality of clustering results, a high Q value means better performance of clustering. A Q value curve gained in the process of hierarchical clustering for personal name "陈坚 (CHEN Jian)" is drawn in Fig. 6. From the picture, we can find a peak when the number of clusters is 7 and the corresponding Q gains a largest value 0.5428, which represents a best clustering result. Such a Q value will be employed as a cluster stopping measure.

5.3 Noise Cluster Removing

Clustering algorithm using the cluster stopping measure Q can determine the number of clusters automatically for each personal name, which gives a better result than a traditional threshold method. Nevertheless, this algorithm will lead to too many noise clusters with only one document. In this paper, we merge noise clusters into their nearest clusters, which can result in an improvement of clustering performance.

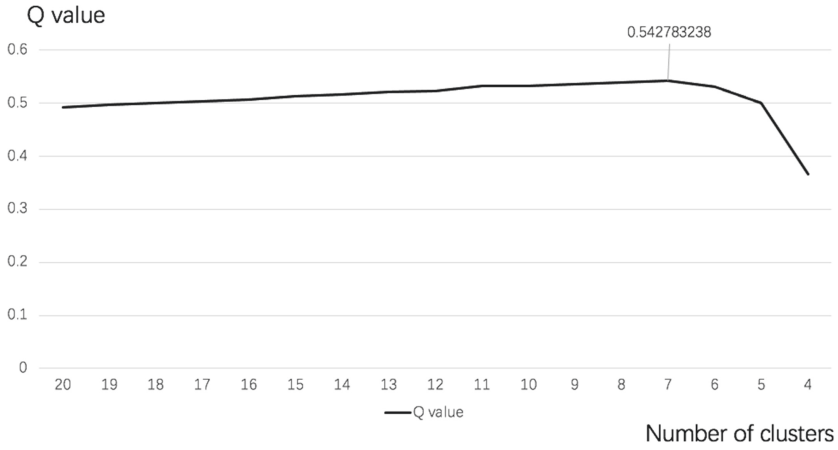

Fig. 6. The Curve of Q value for clustering personal name "CHEN Jian"

5.4 Experimental Result

Experiments are carried out on six groups of personal names for three methods: threshold-based clustering, Q measure-based clustering, and Q measure-based clustering by removing noise clusters. The experimental result for F-score is shown in Fig. 7.

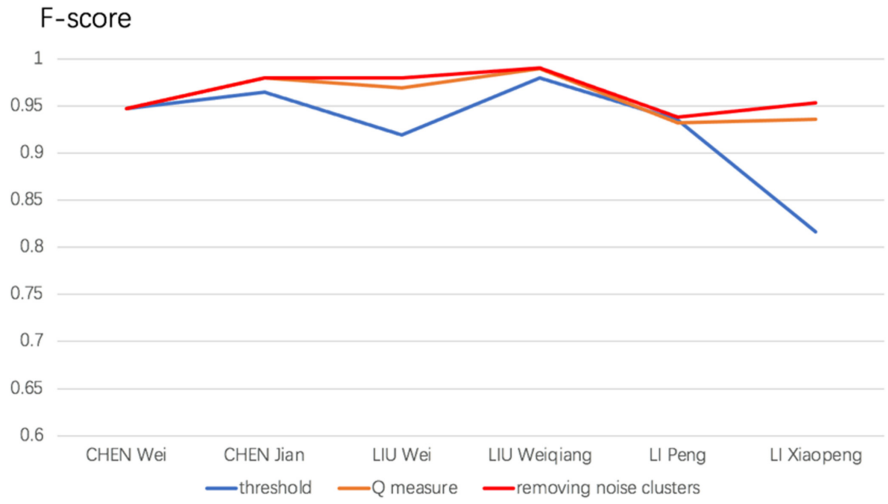

Fig. 7. The F-score of three cluster-stopping algorithm for six personal names. (Color figure online)

As depicted in the Fig. 7, three algorithms achieve the same F-score for personal name "陈卫 (CHEN Wei)". For name "李鹏 (LI Peng)", the threshold-based algorithm

performs better than the Q measure-based algorithm, but worse than the algorithm considering noise cluster removing. Q measure-based algorithm by removing the noise clusters outperforms two other algorithms on average (see Red curve in Fig. 7).

Table 1. Result of noise cluster optimization (# represents the number of clusters)

Personal name	Real number of clusters	# for Q-based method	# for optimization method	F-score for Q-based method	F-score for optimization method
CHEN Wei	5	4	4	94.74%	94.74%
CHEN Jian	7	7	7	98.00%	98.00%
LIU Wei	6	9	7	96.91%	97.96%
LIU Weiqiang	5	7	5	98.99%	99.00%
LI Peng	8	13	10	93.18%	93.83%
LI Xiaopeng	3	9	6	93.62%	95.29%

Table 1 displays the result of cluster optimization experiment for Q measure-based algorithm by removing the noise clusters. Even though the results of two clustering algorithms are the same for "CHEN Wei" and "CHEN Jian", the noise cluster optimization approach performs better for the rest groups. Therefore, the Q measure-based algorithm by removing the noise clusters will reduce the number of clusters and eventually improve the effect of clustering algorithm.

6 Conclusions

This paper has proposed a framework to disambiguate personal names in Chinese dataset. The dataset was created by crawling Baidu news with six personal names and removing HTML tags for standardization. A lot of work of lexical analysis has been done, like segmentation, part-of-speech tagging, and named entity recognition. On one hand, different words were extracted as features and six groups of features were selected for clustering. On the other hand, boolean, absolute frequency, tf-idf, and entropy weights were chosen for the weighting scheme. Finally, twenty-four combinations were tested and feature 5 (noun feature with df $>= 2$) plus tf-idf weights achieved the best result.

As for clustering algorithm, we explored the cluster stopping measure by utilizing a Q measure, which can exhibit better performance than threshold-based method. Moreover, noise cluster optimization was done by merging noise clusters into their neighbor clusters with the largest similarity. Experimental results have shown that the final number of noise clusters is reduced and the proposed approach has a better effect on the dataset of six personal names. However, there are some limitations of the Q measure-based method. The algorithm highly depends on the internal similarity across attributes of documents when stopping the clustering. It inclines to cluster with many small groups.

In future, multiple feature generation will be expanded by introducing syntactic information, which is neglected in our research. Furthermore, the method cannot perform well when the lexical analysis fails, so improving the accuracy of early analysis is a direction of future work.

Acknowledgement. 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. Li, G., Wang, H.: Chinese named entity recognition and disambiguation based on multi-stage clustering. J. Chin. Inf. Process. **27**(5), 29–34 (2013)
2. Zhao, L., Yan, Z., Liang, H.: Ontology-based personal name disambiguation on the web. In: Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), vol. 01 (2013)
3. Emami, H.: A graph-based approach to person name disambiguation in web. ACM Trans. Manag. Inf. Syst. **10**(2), 4.1–4.25 (2019)
4. Xiong, L., et al.: Chinese name disambiguation based on analysis of sentential semantic structure. Appl. Res. Comput. **33**(10), 2898–2901 (2016)
5. Yang, Y., Zhou, J., Li, B.: Name disambiguation algorithm based on ensemble. Appl. Res. Comput. **33**(9), 2716–2720 (2016)
6. Zeng, J., et al.: Research on expert disambiguation of same name based on multi-feature fusion. Acta Scientiarum Naturalium Universitatis Pekinensis **56**(4), 607–613 (2020)
7. Shang, Y., et al.: Co-author and afiliate based name disambiguation approach. Comput. Sci. **45**(11), 220–225 (2018)
8. Zhai, X., et al.: Research on English author name disambiguation based on sparse distributed representation. Appl. Res. Comput. **36**(12), 3534–3538 (2019)
9. Yu, C., et al.: Author name disambiguation with network embedding. Data Anal. Knowl. Disc. **4**(2/3), 48–59 (2020)
10. Pooja, K.M., Mondal, S., Chandra, J.: A graph combination with edge pruning-based approach for author name disambiguation. J. Am. Soc. Inf. Sci. **71**(1), 69–83 (2020)
11. Kim, K., Rohatgi, S., Giles, C.L.: Hybrid deep pairwise classification for author name disambiguation. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp. 2369–2372 (2019)
12. Delgado, D., et al.: Person name disambiguation on the web in a multilingual context. Inf. Sci. **465**, 373–387 (2018)
13. Khabsa, M., Treeratpituk, P., Giles, C.L.: Online person name disambiguation with constraints. In: Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries, Knoxville, TN, pp. 37–46 (2015)
14. Hotho, A., Staab, S., Stumme, G.: WordNet improves text document clustering. In: Proceedings of the SIGIR 2003 Semantic Web Workshop, pp. 541–544 (2003)
15. Newman, M.E.J.: Modularity and community structure in networks. Proc. Natl. Acad. Sci. **103**(23), 8577–8582 (2006)
16. Fan, C., Toriumi, F.: High-modularity network generation model based on the muitilayer network. Trans. Jpn. Soc. Artif. Intell. **32**(6), B-H42_1-11 (2017). https://doi.org/10.1527/tjsai.B-H42