Name Disambiguation in Biendata: Clustering, Heterogeneous graph random walk, and Binary Classification

GuanHua Chen [#1], MingJun Chen [*2]

[#] *Renmin University of China, Department of Information*
59 Zhongguancun Street, Haidian District, Beijing, China
[1] 17chenguanhua@ruc.edu.cn

[*] *Renmin University of China, Department of Information*
59 Zhongguancun Street, Haidian District, Beijing, China
[2] myccmj@163.com

Abstract—**Biendata is an open source data competition community that provides all AI researchers with data sets, open source model code, and a platform for communication and competition. We use the data set in this community to solve the Name Disambiguation problem through three different strategies. Firstly, a novel feature learning method (Triple neural network) is adopted to fine-tune word embedding results to enhance the effect of clustering, and an end-to-end cluster number estimator (LSTM Encoder) is developed to guide the effect of cluster analysis. Secondly, we complete the characterization of the documents by two means with heterogeneous graph random walk and semantic training, thus construct the similarity matrix of the documents. After pre-clustering by clustering algorithm, we further classify the outliers with strong rules. In the third strategy, we transform clustering tasks into a dichotomous problem. First, we directly judge whether the authors of the two papers belong to the same person instead of clustering analysis. Then, we try to use strong rules and similarity matrix for pre-clustering, and then combine the clusters obtained by pre-clustering to get the final results. All of our strategies scored well in Biendata's online rankings, indicating the effectiveness of all three frameworks we provide.**

keywords: Clustering, Heterogeneous Graph Random Walk, Binary Classification

I. INTRODUCTION

Name Disambiguation is a problem that has been studied extensively in various fields for decades. Its practical applications include distinguishing users sharing the same name on social networks, matching records in databases of different enterprises, and distinguishing document authors sharing the same name. In this paper we use data from the Biendata community. Biendata is an open source data competition community that provides all AI researchers with data sets, open source model code, and a platform for communication and competition. Biendata provides us with a processed data, and its online ranking gives us instantly feedback on the effects of the model.

In dealing with the problem of Name Disambiguation, we mainly need to overcome the following problems:

(1) How to represent document entities and calculate the similarity between different entities

(2) How to determine the number of people sharing the same name then complete clustering

(3) How to judge and deal with outliers

To overcome these three problems, we adopt three strategies. In the first strategy, we characterize documents and develop a triple neural network to reduce the dimension of the data. After the dimension reduction, the distance between different clusters in the data is enlarged, which contributes to cluster analysis. We develop an encoder based on LSTM model to predict the number of people sharing the same name, and take the results of the model as the super-parameter of the clustering algorithm.

In the second strategy, we calculate the similarity between documents under the same name to determine clustering and outliers. We characterize the document through heterogeneous graph random walk and semantic training model respectively, and then calculate the similarity matrix. After pre-clustering the similarity matrix, we developed a couple of strong rules to classify the samples judged as outliers during the pre-clustering.

In the third strategy, we transform the clustering problem into a dichotomy problem. For a specific name, we use the SVM dichotomy machine to determine whether each possible pairs of documents belong to the same author. Then we adopt strong rules and similarity matrix to obtain the pre-clustering results, and use SVM technique to combine the small clusters that might be written by the same author into large clusters.

We tested our program on real world data. Our experiment shows that the three strategies have achieved good results on the Biendata's online rankings, ranking top 50.

II. DEFINITION OF PROBLEM

Let a be a given name reference and each document under the author name is named $\mathcal{D}_i^a$, the author identity of each paper is defined as $\mathbb{I}(\mathcal{D}_i^a)$, so if two papers have the same author then $\mathbb{I}(\mathcal{D}_i^a) = \mathbb{I}(\mathcal{D}_j^a)$. Accordingly, we can define the Name Disambiguation task as:

Definition 3.1. Name Disambiguation. The task of author disambiguation is to find a function Φ to partition $\mathcal{D}^a$ into a set of disjoint clusters, i.e.,

$$\Phi(\mathcal{D}^a) \rightarrow \mathcal{C}^a, where\ \mathcal{C}^a = \{\mathcal{C}_1^a, \mathcal{C}_2^a, \dots, \mathcal{C}_K^a\},$$

such that each cluster only contains documents of the same identity i.e., $\mathbb{I}(\mathcal{D}_i^a) = \mathbb{I}(\mathcal{D}_j^a), \forall(\mathcal{D}_i^a, \mathcal{D}_j^a) \in \mathcal{C}_K^a \times \mathcal{C}_K^a$, and different clusters contains documents of different identities i.e., $\mathbb{I}(\mathcal{D}_i^a) \neq \mathbb{I}(\mathcal{D}_j^a), \forall(\mathcal{D}_i^a, \mathcal{D}_j^a) \in \mathcal{C}_K^a \times \mathcal{C}_{K'}^a, k \neq k'$.We define the number of elements in each cluster $\mathcal{C}^a$ to be K.

We define two restrictions: given $\mathcal{D}_i^a$, for any $\mathcal{D}_j^a$ that $\mathbb{I}(\mathcal{D}_i^a) = \mathbb{I}(\mathcal{D}_j^a)$, we define that $\mathcal{D}_j^a$ conforms to the PC rule of $\mathcal{D}_i^a$ (positive restriction). In contrast, given $\mathcal{D}_i^a$, for any $\mathcal{D}_j^a$ meeting $\mathbb{I}(\mathcal{D}_i^a) \neq \mathbb{I}(\mathcal{D}_j^a)$, we say that $\mathcal{D}_j^a$ meets the NC rule of $\mathcal{D}_i^a$ (negative constraint). We will use both of these restriction rules in triplet sampling in our experiment.

We define a random walk path as P and CA rule as P walk along the side of co-authors while CO rule as P walk along the side of co-orgnizations. EOF state is defined as the state in which the next node cannot be searched according to CA or CO rule, or when the number of walking length reaches the given limit. EOF means the stop of a random walk.

III. FRAME WORK

Strategy I: Clustering:

In this strategy, we adjust the word embedding results with a triple neural network, and use the results of the cluster number estimator to guide the classification algorithm.

Feature selection and pre-processing

We delete non-English characters, stop words and words with word length less than 3. Then we used Word2vec model to train the corpus. We verify the effect of the training results by printing out a list of words which are most similar to a given word, and observe the degree of similarity between them.

Global Metric Learning

The results of word embedding are still limited to distinguish different authors, so we developed a network with supervised information to fine-tune word embedding vectors.

We use the feedforward neural network using Triplet Loss as its Loss function to make the Euclidean distance between documents in the same cluster smaller and the Euclidean distance between documents in different clusters larger. This technique can effectively enhance the effect of clustering.

Triplet Loss

Let (D_i, D_{i+}, D_{i-}) be a triplet where $\mathbb{I}(D_i) = \mathbb{I}(D_{i+})$ and $\mathbb{I}(D_i) \neq \mathbb{I}(D_{i-})$, we have

$$\delta(y_i, y_{i+}) + m < \delta(y_i, y_{i-}),$$
$$\forall(D_i, D_{i+}, D_{i-}) \in \tau$$

where τ is the set of all possible triplets in the training set, m is a margin enforced between positive pairs and negative pairs, The objective $\mathcal{L}_f$ is then

$$\mathcal{L}_f = \sum_{(D_i, D_{i+}, D_{i-}) \in \tau} \max\{0, \delta(y_i, y_{i+}) - \delta(y_i, y_{i-}) + m\}.$$

Instead of projecting to a single point, triplet loss enables documents with the same identity to reside on a manifold, and at the same time maintain a distance from other documents.

The structure of the network is shown in Figure 1.

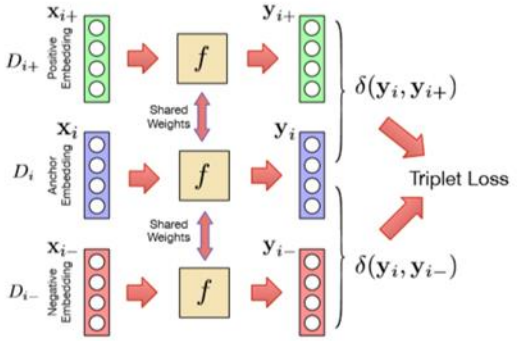

Fig. 1. The structure of the network

To properly construct the triplet training set. For a data set $\mathcal{D}^a$ and its corresponding $\mathcal{C}^a$, for each $\mathcal{D}_i^a$ in every $\mathcal{C}_i^a$ we randomly select six P_j(j=1,2,3,4,5,6) that meet the PC rule of $\mathcal{D}_i^a$, and six $\mathrm{N_j}$($\mathrm{N_j} \in \mathcal{C}^a - \mathcal{C}_i^a$, j=1,2,3,4,5,6) that meet the NC rule of $\mathcal{D}_i^a$. Finally, we get six triplets consisting of one sample, 6 positive samples and 6 negative samples ($\mathcal{D}_i^a, P_j, \mathrm{N_j}, j$=1,2,3,4,5,6). The advantage of this sampling method is to avoid the large training set and balance the data distribution of the training set on each cluster.

Cluster Size Estimation

Cluster analysis based on Global Metric Learning is still not suitable for clustering the number of super-par for lacking super-parameter cluster number. Due to the uncertainty of Euclidean distance between documents within each cluster, density-based algorithms like DBSCAN cannot achieve good results.

We use an RNN encoder and try to map vectors encoded by word embedding to a correct number of clusters. In order to achieve this goal, the main challenge is that the length of candidate sets vary in range from one to hundreds and thousands. Although the RNN neural network can process variable-length data, it still has bias.

To solve this problem, we adopt a sampling strategy to construct a pseudo-training set. Let $c^a = \{c_1^a, c_2^a, \dots\}$ be a collection of clusters and the author identity of documents in each cluster is the same. For each training step, we randomly select a k_t from [5, K] (K refers to the number of clusters in c^a) for c^a. Then we randomly select k_t clusters from K clusters to construct a set C_t. let Dc_t represent documents in C_t. We sample 300 documents from Dc_t with replacement. In this way we can construct infinite training sets from the given data. We use a bidirectional LSTM neural network as the encoder and a fully connected neural network as the decoder. The model takes the result of word embedding as input. We take the mean square error of the logarithm as the loss function, as shown in the formula below.

$$\mathcal{L}_h = \frac{1}{N}\sum_{t=1}^{a}[\log(1 + h(\mathcal{D}_t)) - \log(1 + \mathrm{K})]^2$$

Strategy II: Similarity Matrix Based on Heterogeneous Graph Random Walk

In this strategy, semantic training and heterogeneous graph random walk are used to represent a document, and then the similarity matrix of documents under each name is calculated. We obtain pre-clustering results through clustering analysis. For those documents judged as outliers in the pre-clustering process, we set some strong rules to classifier them and get the final clustering result.

Constructing of heterogeneous network

A heterogeneous graph is a graph in which there is more than one measure of the relationship between two nodes. Random walk refers to the process in which nodes in the graph move to other nodes according to certain rules. This process is martensitic and will forget the history information. Heterogeneous graph and random walk are important tools in representation learning and feature engineering, and they are widely used in Internet link analysis and financial stock market analysis.

Deciding relationships between nodes is a challenge. We choose the co-authors of the paper and orgnization word information to measure the relationship between nodes.A CoAuthor edge between two nodes means a co-author exists both in two papers (not including the name that needs to be disambiguated), while a CoOrg edge between two nodes means a same word exists in the name of the organization to which the authors of two papers belong, the heterogeneous graph we design is an unweighted heterogeneous graph network.

Random walk based on meta-path

We construct the meta-path of document ID by randomly walking on the heterogeneous graph, taking the path as the input of word embedding model to finally get the representation of each paper. Specifically, each node in the network is taken as the initial node in turn. We decide that in the process of random walk, CA and CO take turns to be the rule of random walk, and the target of each walk is randomly selected. We determine the upper limit of the length of a meta path in a random walk in advance, and set P to enter the STATE of EOF when the upper limit is reached or the next node cannot be found according to the current rule. The random walk path is stored in a file for word embedding. The results of the training will be used to characterize the document.

Calculation of similarity matrix

We respectively use the semantic training and the heterogeneous graph random walk to characterize a document in two ways (the semantic training uses the results of word embedding in Strategy one). We calculate the arithmetic mean of the cosine of the Angle between the two documents calculated by two representation methods as an index to measure the similarity of the two documents.

We construct the similarity matrix for all papers under each name and uase clustering analysis to get the result of pre-clustering.

Strong rules clustering for outliers

Strong rule clustering means to classify samples using a series of external rules. Here, we classify the outliers by quantifying the strength of the relationship between them and other documents, and classify them into clusters having the strongest relationship with them when their relationship strength exceeds a threshold. The strength of the relationship between a paper and a cluster is replaced by the strength of the relationship between the paper and the paper with the strongest relationship with it in the cluster. Factors taken into account include the number of the same co-authors, the number of the same institution name words and the degree to which paper titles and keywords are similar.

Strategy III: Binary classification

In Strategy three, we transform the classification task into a dichotomy problem.

Feature extraction

By extracting the common features of the two papers, we use vectors to describe their similarity. We select the number of same co-authors, the number of repeated words in the name of the author's organization and the number of repeated words in the title and keyword of two papers to form the vector.

Direct dichotomy

For each author name, we use a SVM classifier on every possible paper pairs to determine whether they belong to the same author.

Pre-clustering + dichotomy

We use strong rules and similarity matrix to pre-cluster papers, getting many small clusters. Then we define the similarity between the two clusters with the similarity of the two papers that are most similar in the two clusters. If the similarity between the clusters exceeds the threshold, we will combine these two clusters together. The strong rule is the same as that in strategy two. The similarity matrix is calculated by the vector extracted in the feature extraction step.

IV. Discussion

Global Metric Learning is instructive because it is also applicable to papers that do not appear in the training process, effectively transferring the known supervisory information to unknown situations. Meanwhile, based on the appropriate sampling method, the training speed of Global Metric Learning is relatively fast. Global Metric Learning can effectively fine-tune the results of word embedding and optimize the results of clustering.

Cluster Size Estimation is instructive for Cluster analysis in most cases. However, since the loss function takes the logarithm before calculating the mean square error, the difference between the predicted value and the actual value will still be too large when the model fits well. It is more suitable for the data set with a more balanced number of clusters. With the increase of the dispersion degree of the number of clusters, the overall effect of Cluster Size Estimation will vary greatly.

The effectiveness of heterogeneous graph random walk has been verified in various known results, by constructing the meta path papers are well characterized. In addition, heterogeneous graph random walk has another advantage that it does not require training set and prior knowledge. The method of using strong rule clustering to classify outliers can be used as a

supplement to the pre-clustering, which effectively improves the accuracy of prediction.

The dichotomy strategy is simple to understand and easy to implement, but it is more dependent on the selection of features and super parameters, and requires more experience of programmers. Direct dichotomy method risks incorporating two wrong large cluster, which is deadly harmful to the prediction result. Although strong rule clustering can bring the benefits of prior knowledge to binary classification model, the effect is not stable considering that the data in the real world is very complicated. The combination of similarity matrix and the binary classification taking advantages from both strategy two and three, overcoming the lack of stability. Due to the large amount of data to be calculated, strategy three may not be suitable for the instant updating task of online platform.

V. Experiment

Environment

Experimental platform: Windows 10
Programming language: python3.6
Neural network building tool: Pytorch 1.4.0+ CPU
CPU model: Intel Core I7-8750h

Data Overview

The training set is divided into author data and pub data. The author data stores the paper ID under the person's name in JSON format and the dictionary key is 101 person names. The value is the author ID sharing that name and the list of paper ids written by that ID, represented as a dictionary. Pub data in json format stores the information of 120000 papers, the keys of the dictionary is paper id, value is an information dictionary, the keys of the dictionary is the title, abstract, key words, author, journal name and the publication year of the paper. The corresponding value to key "author" is a dictionary list, each of the element in the list contains the name of the author and the organization he belongs to.

Process of Experiment

We set the output dimension of Word2vec as 100 and choose to print the list of words most similar to the word ‘Simple’ to verify the effect of model training. The list of words is shown in Table 1.

TABLE I

The word most similar to ‘simple’	similarity	rank
Convenient	0.86	1
Easy	0.79	2
Feasible	0.77	3
Developed	0.77	4
Straightforward	0.75	5

In strategy one, we set two hidden layers in the Triplet Loss neural network, set the number of hidden nodes as 100,64 and set the activation function as ReLu. The Margin in Triplet Loss is set 1. In the cluster number estimator, we use a bidirectional LSTM layer whose output dimension is 128 and a fully connected neural network to map the input to a constant.

In strategy two, we take the result of word2vec in strategy one as the result of semantic characterization. In the random walk of heterogeneous graph, we set the maximum length of per walk to be 20, and perform 5 traversals of the whole graph. Each traversal takes each node in the graph as the starting one of a random walk in turn.

In the setting of strong rule clustering, we calculate the similarity value of two papers by formula 1.5* duplicate author number +1* duplicate words in institution name +1/3* duplicate words in title and keyword. Let 1.5 be the threshold of the relational value.

In strategy three, gaussian kernel function is selected as the kernel function of SVM classifier. The formula of feature selection and strong rule formulation is the same as that in strategy two. The threshold of similarity between two clusters is 0.76, beyond which we merge the two clusters.

Results of Experiment

The loss function of Global Metric Learning is shown in Figure 2 (the loss value is printed once for every 1000 training step)

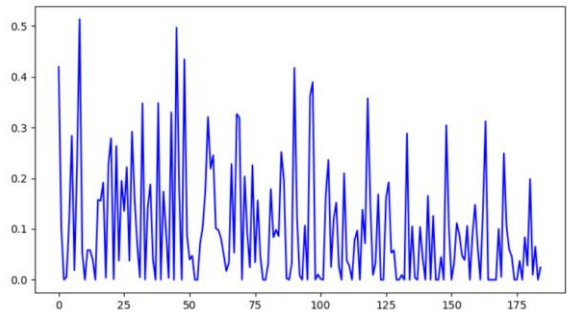

Fig. 2.

It can be seen that the loss value is decreasing in general. Considering that our dynamic sampling strategy will increase the uncertainty of data, the fluctuation of the loss function can be explained.

The training loss function of Cluster Size Estimation is shown in Figure 3

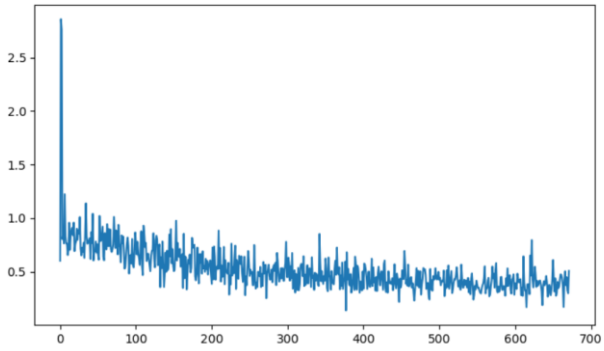

Fig. 3.

It can be seen that the training is effective.

Table 2 shows the performance of the three strategies on the online rankings in Biendata. The site uses F1 values as a measure of prediction.

TABLE 2

Strategy	F1
Word Embedding+ Cluster Number Estimator	0.47
Word Embedding+ Global Metric Learning + Cluster Number Estimator	0.49
Direct Binary Classification	0.53
Strong rules + Binary Classification	0.71~0.85
Heterogeneous Graph Random walk	0.80
Similarity Matrix+ Binary Classification	0.89

After research, we find that the effect of strategy one is generally unsatisfactory because it is weak to identify outliers. More importantly, there is a large gap between the predicted value by the Cluster Number Estimator and the true value (as is mentioned in Discussion section) when the number of clusters is too large.

In addition, the effect of using Global Metric Learning is about 2% higher than not using it, which verifies the effectiveness of Global Metric Learning on fine-tuning word embedding results.

Compared with strategy one, strategy two is better at distinguishing outliers and constructing similarity matrix with comprehensive method is also beneficial to cluster analysis. On the online ranking, the performance of strategy two is better than strategy one.

In the experiment of strategy 3, direct binary may mistakenly combine two large clusters together, seriously damaging the prediction accuracy, so it performs only moderately. Strong rule clustering helps classification from the perspective of prior knowledge, so the effect is obviously better. However, due to the complexity of real data, the performance of this scheme is unstable, with F1 value fluctuating between 0.71 and 0.85. The scheme combining similarity matrix with dichotomy has the best performance on the test data and is stable at 0.89 because it combines the core ideas of strategy two and strategy three. It is the model with the best performance in this paper.

VI. Conclusion

In strategy one, Global Metric Learning has been proved to be effective. Although the Estimator of Cluster Number has poor prediction ability in extreme cases, it is suitable for the situation where the number of clusters is relatively more balanced. Strategy two uses an appropriate method to construct the similarity matrix between papers. The DBSCAN pre-clustering method has advantages in distinguishing outliers in nature and the strong regular clustering is a good supplement to the pre-clustering. Strategy three generally adopts the idea of dichotomy. After continuous attempts, it is found that combining with the idea of strategy two, on the basis of pre-clustering completed by using similarity matrix, the scheme of combining small clusters into large clusters by using dichotomy can obtain excellent and stable results.

ACKNOWLEDGMENTS

Thanks to the course of Unstructured Big Data Analysis, we can have the opportunity to study the problem of name disambiguation. In this paper, we adopt various methods and provide three effective solutions to this problem. At the same time, I would also like to thank Ms. Zhang Jing for her teaching during the semester. Despite the epidemic and difficulties in online teaching, she still ensured the quality of the course, making sure that every student is making progress. We express gratitude to TA Tang Xiaobin, who provided us with patient guidance during our research.

References

[1] S. M. Metev and V. P. Veiko, *Laser Assisted Microtechnology*, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.

[2] J. Breckling, Ed., *The Analysis of Directional Time Series: Applications to Wind Speed and Direction*, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, vol. 61.

[3] S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A novel ultrathin elevated channel low-temperature poly-Si TFT," *IEEE Electron Device Lett.*, vol. 20, pp. 569–571, Nov. 1999.

[4] M. Wegmuller, J. P. von der Weid, P. Oberson, and N. Gisin, "High resolution fiber distributed measurements with coherent OFDR," in *Proc. ECOC'00*, 2000, paper 11.3.4, p. 109.

[5] R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, "High-speed digital-to-RF converter," U.S. Patent 5 668 842, Sep. 16, 1997.

[6] (2007) The IEEE website. [Online]. Available: http://www.ieee.org/

[7] M. Shell. (2007) IEEEtran webpage on CTAN. [Online]. Available: http://www.ctan.org/tex-archive/macros/latex/contrib/IEEEtran/

[8] *FLEXChip Signal Processor (MC68175/D)*, Motorola, 1996.

[9] "PDCA12-70 data sheet," Opto Speed SA, Mezzovico, Switzerland.

[10] A. Karnik, "Performance of TCP congestion control with rate feedback: TCP/ABR and rate adaptive TCP/IP," M. Eng. thesis, Indian Institute of Science, Bangalore, India, Jan. 1999.

[11] J. Padhye, V. Firoiu, and D. Towsley, "A stochastic model of TCP Reno congestion avoidance and control," Univ. of Massachusetts, Amherst, MA, CMPSCI Tech. Rep. 99-02, 1999.

[12] *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification*, IEEE Std. 802.11, 1997