Computer Engineering and Applications Vol. 8, No. 1, February 2019

Deep Neural Network Structure to Improve Individual Performance based Author Classification

Firdaus, Muhammad Anshori, Sarifah Putri Raflesia, Mira Afrina, Ahmad Zarkasi, Siti Nurmaini

Intelligent System Research Group, Faculty of Computer Science, Universitas Sriwijaya, Indonesia
virdauz@gmail.com, msstoci@gmail.com, syarifahpr@gmail.com, mira_afr@yahoo.com.my, zarkasi98@gmail.com, sitinurmaini@gmail.com

ABSTRACT

This paper proposed an improved method for author name disambiguation problem, both homonym and synonym. We applied Deep Neural Network by using a self-developed dataset extracted from Scopus. The dataset consists of 6 attributes. The data prepared for the DNN is the distance data of each pair of authors' attributes, Levenshtein distance are used. Using DNN, we found large gains on performance. The result shows the level of accuracy is 99.6% on average.

Keywords: Author Name Disambiguation, Bibliographic Repository, Deep Neural Network.

1. INTRODUCTION

Scientific digital libraries have become an important source of bibliographic notes for the scientific community[1]. It becomes an important source for scientists to get literature and find interesting topics[2]. It also provide analysis that can be used for better decision making by donors and academic institutions to determine grant recipients and individual promotions[3]. With the growing size of scientific digital libraries, it becomes a challenge to identify the author correctly and put the publication to them [4].

The quality of scientific digital library data depends on the Author Name Disambiguation (AND) process, which links the author's name to the right person[5]. AND is an important issue that needs to be solved in bibliometric analysis of a scientific digital library [2]. AND may occur due to multiple authors with the same name (homonym) or different name variations for the same person (synonym) [6][4].

Several techniques have been proposed to solve AND problems. In general, techniques for solving AND problems are divided into two groups [3]; 1) machine learning-based techniques consists of supervised [7]–[11], unsupervised [12]–[16] and semi-supervised [17]–[25], 2) non machine learning-based techniques consists of graph-based [26][27][28][29][30][31] and heuristic-based techniques [32][33]. Machine learning techniques build models based on previous observations [34]. It used to predict the class of data that is not visible[34].

In AND solution, the supervised technique has better performance on its training dataset than any other technique, but it requires a lot of training data for each class. It requires complete training data that represents each of its target classes. Whereas in

unsupervised techniques, the selection of similarity sizes and clustering techniques to match, becomes a difficult task. The semi-supervised technique works well when the ambiguous target author is limited, but fails as the number increases. Graph-based techniques are the new techniques applied to the AND solution, so it still needs more convincing verification. While heuristic techniques have not produced stable results.

Deep Neural Network (DNN) is one of the supervised machine learning techniques to solve AND problems. DNN is an example of Deep architecture model, Deep learning itself is a concept of Artificial Neural Network that is in between the input and output layer. The advantages of this method is to extract features by learning data and not forcing feature based on pre-processing results like filtering and thresholding[35].

To the best of my knowledge, there is only one study that uses DNN to solve the AND problem. Using the Vietnamese author dataset, Tran produces a good degree of accuracy [10]. In this paper we use DNN to solve AND problems by using a self-developed dataset extracted from Scopus.

2. METHOD

2.1. DATA PREPARATION

There are various datasets that have been used, both from Digital Library extraction and synthetic dataset. Digital libraries used as sources are as follows; 1) Brazilian Digital Library of Computing (BDBComp) [12][17][36][12][37][38], 2) Digital Bibliography and Library Project (DBLP) [4][9][12][22][17][13][15][39][40][19][36][26][27][41][42][37][20][25][43][7], 3) Arnetminer [13][15][27], 4) Microsoft Academic Search (MAS) [44][32][24][40], 5) CiteSeer [15][40]. Some studies use self-developed datasets, such as Chinese, Korean, German and Vietnamese datasets.

TABLE 1.
Author Name Disambiguation Dataset Description

Author	Publication Number
Sidik, M. A. B	52
Firdaus, F	3
Firdaus	4
Firdaus, M	2
Nurmaini, S	38
Saparudin	15
Setiawan, B	5
Setiawan, B	5
Setiawan, B	1

The dataset used in this research is self-developed data extracted from Scopus. Data is extracted and labeled manually. The dataset consists of 125 publications from 9 Indonesian authors with the following attributes; author name, title, year, source, author affiliation, co-authors name (Table 1). The dataset contains author homonym

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

and synonym data. For each authors and publications are given a unique identification number.

2.2 DATA PREPROCESSING

For each row of publication data is paired with another. Paired data with the same author is labelled with 1 and 0 for different author. There are 7750 publication data pairs, 2202 same author and 5548 different author pairs.

The data prepared for processing in DNN is the distance data of each pair of attributes, in this paper we used Levenshtein distance (Equation 1).

$$lev_{a,b}(i,j) = \begin{cases} \max(i,j) & if\ \min(i,j) = 0, \\ min \begin{cases} lev_{a,b}(i-1,j)+1 \\ lev_{a,b}(i,j-1)+1 \\ lev_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} & otherwise. \end{cases} \quad (1)$$

For all attributes, the values is rescaled so that they have the properties of a standard normal distribution. In this case we used Z-score normalization, the data is scaled to 0 to 1 range (Equation 2).

$$X_{norm} = \frac{X - X_{min}}{X_{min} - X_{max}} \quad (2)$$

2.3 EXPERIMENTAL SETUP

The experiments conducted in 4 scenarios. Each scenarios using four activation function combination (Table 2).

TABLE 2.
Experiment Scnearios

	Activation Function			
	Experiment 1	Experiment 2	Experiment 3	Experiment 4
Hiden layer	Relu	Relu	Softsign	Softsign
Output layer	Sigmoid	Softmax	Sigmoid	Softmax

For each scenarios, we used 1 to 15 hidden layers, 50 neuron for each layer, 100 epoch, 0.0001 learning rate and binary crossentropy as a loss function. We performed a stratified 10-fold cross-validation to evaluate each of scenarios.

3. RESULTS AND DISCUSSION

Out of the many trials from 4 experiment scenarios, we found various characteristics of results (Figure 1). The structure of the first experiment produces a relatively stable accuracy for each number of layers. The second experiment resulted prospective accuracy in each layer addition, contrary to the third and fourth.

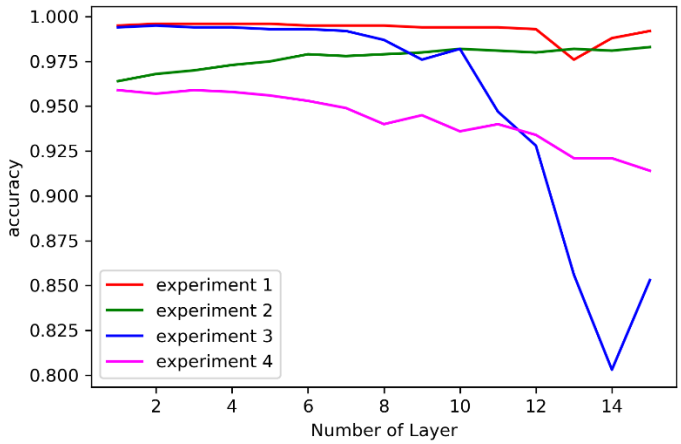

FIGURE 1. Neural Network Structures Accuracy Comparison

Of all the structures tested, structures with Relu and Sigmoid activation functions in the hidden and output layers with the number of layers 2, 3, 4 and 5 (Table 3) produce the highest mean accuracy, which is 99.6%. This structure also obtained an outstanding score in precision, recall and F1 score in sequence 99.5%, 99.7% and 99.6% on average. The accuracy is better than Tran [10].

TABLE 3.
Highest Accuracy Neural Network Structure for AND

Layers	Number of Neuron	Activation Function
Input Layer	6	-
Hidden Layer 1	100	Relu
Hidden Layer 2	100	Relu
Hidden Layer 3	100	Relu
Hidden Layer 4	100	Relu
Hidden Layer 5	100	Relu
Output Layers	1	Sigmoid

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

4. CONCLUSSION

In this paper e new method has been proposed in improving AND solving problems using the Deep Neural Network technique. Of the various structures that were tried, the structure obtained resulted in very good accuracy. In the future, we will use another dataset to get the most suitable structure for solving AND problems in general.

ACKNOWLEDGMENTS

This work was supported by Universitas Sriwjaya which providing financial research Hibah Penelitian Dosen Muda SATEKs 2018 contract No. 0008/UN9/SK.LP2M.PT/2018

REFERENCES

[1] R. Hazra, A. Saha, S. B. Deb, and D. Mitra, "An efficient technique for author name disambiguation," *2016 IEEE Int. Conf. Curr. Trends Adv. Comput. ICCTAC 2016*, 2016.

[2] H. Han, C. Yao, and Y. Fu, "Semantic fingerprints-based author name disambiguation in Chinese documents," *Scientometrics*, 2017.

[3] I. Hussain and S. Asghar, "A survey of author name disambiguation techniques: 2010--2016," *Knowl. Eng. Rev.*, vol. 32, 2017.

[4] F. Momeni and P. Mayr, "Using Co-authorship Networks for Author Name Disambiguation," *Proc. 16th ACM/IEEE-CS Jt. Conf. Digit. Libr. - JCDL '16*, pp. 261–262, 2016.

[5] J. Schulz, "Using Monte Carlo simulations to assess the impact of author name disambiguation quality on different bibliometric analyses," *Scientometrics*, 2016.

[6] X. Lin, J. Zhu, Y. T. B, F. Yang, B. Peng, and W. Li, "A Novel Approach for Author Name," pp. 169–182, 2017.

[7] D. Han, S. Liu, Y. Hu, B. Wang, and Y. Sun, "ELM-based name disambiguation in bibliography," *World Wide Web*, vol. 18, no. 2, pp. 253–263, 2015.

[8] T. Huynh, K. Hoang, T. Do, and D. Huynh, "Vietnamese author name disambiguation for integrating publications from heterogeneous sources," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2013, vol. 7802 LNAI, no. PART 1, pp. 226–235.

[9] N. Onodera *et al.*, "A method for eliminating articles by homonymous authors from the large number of articles retrieved by author search," *J. Assoc. Inf. Sci. Technol.*, vol. 62, no. 4, pp. 677–690, 2011.

[10] H. N. Tran, T. Huynh, and T. Do, "Author name disambiguation by using deep neural network," in *Asian Conference on Intelligent Information and Database Systems*, 2014, pp. 123–132.

[11] J. Wang, K. Berzins, D. Hicks, J. Melkers, F. Xiao, and D. Pinheiro, "A boosted-trees method for name disambiguation," *Scientometrics*, vol. 93, no. 2, pp. 391–411, 2012.
[12] A. P. de Carvalho, A. A. Ferreira, A. H. F. Laender, and M. A. Gonçalves, "Incremental unsupervised name disambiguation in cleaned digital libraries," *J. Inf. Data Manag.*, vol. 2, no. 3, p. 289, 2011.
[13] Y. Liu, W. Li, Z. Huang, and Q. Fang, "A fast method based on multiple clustering for name disambiguation in bibliographic citations," *J. Assoc. Inf. Sci. Technol.*, vol. 66, no. 3, pp. 634–644, 2015.
[14] C. Schulz, A. Mazloumian, A. M. Petersen, O. Penner, and D. Helbing, "Exploiting citation networks for large-scale author name disambiguation," pp. 1–14, 2014.
[15] J. Tang, A. C. M. Fong, B. Wang, and J. Zhang, "A unified probabilistic framework for name disambiguation in digital library," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 6, pp. 975–987, 2012.
[16] L. Tang and J. P. Walsh, "Bibliometric fingerprints: name disambiguation based on approximate structure equivalence of cognitive maps," *Scientometrics*, vol. 84, no. 3, pp. 763–784, 2010.
[17] A. A. Ferreira, A. Veloso, M. A. Gonçalves, and A. H. F. Laender, "Self-Training Author Name Disambiguation for Information Scarce Scenarios," vol. 65, no. 6, pp. 1257–1278, 2014.
[18] T. Gurney, E. Horlings, and P. Van Den Besselaar, "Author disambiguation using multi-aspect similarity indicators," *Scientometrics*, vol. 91, no. 2, pp. 435–449, 2012.
[19] M. Imran, S. Z. H. Gillani, and M. Marchese, "A real-time heuristic-based unsupervised method for name disambiguation in digital libraries," *D-Lib Mag.*, vol. 19, no. 9/10, 2013.
[20] M. Levin, S. Krawczyk, S. Bethard, and D. Jurafsky, "Citation-based bootstrapping for large-scale author disambiguation," *J. Assoc. Inf. Sci. Technol.*, vol. 63, no. 5, pp. 1030–1047, 2012.
[21] G. Louppe, H. T. Al-Natsheh, M. Susik, and E. J. Maguire, "Ethnicity sensitive author disambiguation using semi-supervised learning," in *International Conference on Knowledge Engineering and the Semantic Web*, 2016, pp. 272–287.
[22] H.-T. Peng, C.-Y. Lu, W. Hsu, and J.-M. Ho, "Disambiguating authors in citations on the web and authorship correlations," *Expert Syst. Appl.*, vol. 39, no. 12, pp. 10521–10532, 2012.
[23] P. Wang, J. Zhao, K. Huang, and B. Xu, "A unified semi-supervised framework for author disambiguation in academic social network," in *International Conference on Database and Expert Systems Applications*, 2014, pp. 1–16.
[24] J. Zhao, P. Wang, and K. Huang, "A semi-supervised approach for author disambiguation in KDD CUP 2013," in *Proceedings of the 2013 KDD Cup 2013 Workshop*, 2013, p. 10.
[25] J. Zhu, Y. Yang, Q. Xie, L. Wang, and S.-U. Hassan, "Robust hybrid name disambiguation framework for large databases," *Scientometrics*, vol. 98, no. 3, pp. 2255–2274, 2014.
[26] B.-W. On, I. Lee, and D. Lee, "Scalable clustering methods for the name disambiguation problem," *Knowl. Inf. Syst.*, vol. 31, no. 1, p. 129, 2012.
[27] D. Shin, T. Kim, J. Choi, and J. Kim, "Author name disambiguation using a graph model with node splitting and merging based on bibliographic

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

information," *Scientometrics*, vol. 100, no. 1, pp. 15–50, 2014.
[28] X. Fan, J. Wang, X. Pu, L. Zhou, and B. Lv, "On graph-based name disambiguation," *J. Data Inf. Qual.*, vol. 2, no. 2, p. 10, 2011.
[29] F. H. Levin and C. A. Heuser, "Evaluating the use of social networks in author name disambiguation in digital libraries," *J. Inf. Data Manag.*, vol. 1, no. 2, p. 183, 2010.
[30] X. Wang, J. Tang, H. Cheng, and S. Y. Philip, "Adana: Active name disambiguation," in *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, 2011, pp. 794–803.
[31] Y. Liu and Y. Tang, "Network based Framework for Author Name Disambiguation Applications," *Int. J. u-and e-Service, Sci. Technol.*, vol. 8, no. 9, pp. 75–82, 2015.
[32] W.-S. Chin *et al.*, "Effective string processing and matching for author disambiguation," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3037–3064, 2014.
[33] A. F. Santana, M. A. Gonçalves, A. H. F. Laender, and A. A. Ferreira, "On the combination of domain-specific heuristics for author name disambiguation: the nearest cluster method," *Int. J. Digit. Libr.*, vol. 16, no. 3–4, pp. 229–246, 2015.
[34] F. Provost and R. Kohavi, "Guest editors' introduction: On applied research in machine learning," *Mach. Learn.*, vol. 30, no. 2, pp. 127–132, 1998.
[35] J. Masci, A. Giusti, D. Ciresan, G. Fricout, and J. Schmidhuber, "A fast learning algorithm for image segmentation with max-pooling convolutional networks," in *2013 IEEE International Conference on Image Processing*, 2013, pp. 2713–2717.
[36] A. Veloso, A. A. Ferreira, M. A. Gonçalves, A. H. F. Laender, and W. M. Jr, "Cost-effective on-demand associative author name disambiguation," *Inf. Process. Manag.*, vol. 48, no. 4, pp. 680–697, 2012.
[37] A. F. Santana, M. Andr, A. H. F. Laender, and A. A. Ferreira, "Incremental Author Name Disambiguation by Exploiting Domain-Specific Heuristics," vol. 00, no. 00, 2016.
[38] A. A. Ferreira and M. A. Gonçalves, "A Brief Survey of Automatic Methods for Author Name Disambiguation," vol. 41, no. 2, 2012.
[39] N. R. Smalheiser and V. I. Torvik, "Author name disambiguation," *Annu. Rev. Inf. Sci. Technol.*, vol. 43, no. 1, pp. 1–43, 2009.
[40] P. Andruszkiewicz and S. Szepietowski, "Person Name Disambiguation for Building University Knowledge Base," no. Ii, pp. 270–279, 2016.
[41] A. A. Ferreira, R. Silva, M. A. Gonçalves, A. Veloso, and A. H. F. Laender, "Active Associative Sampling for Author Name Disambiguation," pp. 175–184, 2012.
[42] H. Peng, C. Lu, W. Hsu, and J. Ho, "Expert Systems with Applications Disambiguating authors in citations on the web and authorship correlations," *Expert Syst. Appl.*, vol. 39, no. 12, pp. 10521–10532, 2012.
[43] K. Yang and Y. Wu, "Author Name Disambiguation in Citations," pp. 335–338, 2011.
[44] J. Liu, "Ranking-Based Name Matching for Author Disambiguation in Bibliographic Data."