Improving Academic Homepage Identification from the Web Using Neural Networks

Jiapeng Zhao[1,2], Tingwen Liu[1,2(✉)], and Jinqiao Shi[1,3]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
{zhaojiapeng,liutingwen,shijinqiao}@iie.ac.cn
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3] Key Lab of Trustworthy Distributed Computing and Service (BUPT), Ministry of Education, Beijing, China

Abstract. Identifying academic homepages is a fundamental work of many tasks, such as expert finding, researcher profile extraction and homonym researcher disambiguation. Many works have been proposed to obtain researcher homepages using search engines. These methods only extract features at the lexical-level from each single retrieval result, which is not enough to identify homepage from retrieval results with high similarity. To address this problem, we first make deep-insight improvements on three aspects. (1) Fine-gained features are designed to efficiently detect whether the researcher's name appears in retrieval results; (2) Establishing correlation of multiple retrieval results for the same researcher; (3) Obtaining semantic information involved in URL, title and snippet of each retrieval result by recurrent neural networks. Afterwards, we employ a joint neural network framework which is able to make comprehensive use of these informative information. In comparison with previous work, our approach gives a substantial increase of 10%–11% accuracy on a real-world dataset provided by AMiner. Experimental results demonstrate the effectiveness of our method.

Keywords: Academic homepage identification · Retrieval results · Semantic information representation · Joint neural network

1 Introduction

The academic homepage of a researcher usually contains lots of profile information and descriptions, such as employment status, research interests, contact information and publications. These are essential resources for the digital library access portals [8] to collect the researcher's metadata. In general, there are two frequently-used ways to collect academic homepages. One is to monitor the official websites of known research institutes and make a binary classification on each crawled webpage to determine whether it is an academic homepage. The other

J. M. F. Rodrigues et al. (Eds.): ICCS 2019, LNCS 11540, pp. 551–558, 2019.
https://doi.org/10.1007/978-3-030-22750-0_50

is to use researcher names and some additional information (such as researcher's affiliation and research interests) as search engine queries to retrieve related webpages from the web, and choose one retrieval result as the academic homepage. In this paper, we focus on the second way to collect academic homepages massively only with retrieval results of search engines including URLs, titles and snippets, because it can be deployed as an API service and respond rapidly with low resource cost. However, building an accurate module that automatically identifies researcher homepages from retrieval results is not easy, owing to the following technical challenges. First, one researcher may have multiple webpages associated with him/her. Second, search engines may split the query into multiple fragments to obtain more retrieve results, but introduce more noise at the same time. Then one's homepage may rank very low in retrieval results.

In this paper, we focus on the academic homepage identification. In comparison with previous work that only extracted some simple statistical features from each retrieval result, Our key contributions are as follows: (1) We proposed a novel solution to identify researcher homepages via search engines, and demonstrated the effectiveness of our approach on a publicly-available dataset. It not only obtains remarkable improvements with respect to the accuracy, but also performs more stable through computing precision and recall by selecting different proportions of test results. (2) We designed four types of novel features to help identify homepage from high similarity retrieval results. (3) We presented a joint neural network model, which allows different kinds of neural networks being trained synchronously, and thus makes full use of hand-crafted features information and sequence information.

2 Related Work

Relevant work on academic homepage identification using retrieval results of search engines first appeared in TREC's track [1]. It's an entity-oriented web search task. The task aims at finding homepages for four types of entities: organization, location, person, and product. To identify an academic homepage, many query-dependent features can be effectively utilized. Tang et al. [10] used researcher's name and affiliation name as queries of search engines and selected the best retrieval results as researcher homepage, but only hand-craft features from URL are used in their work.

In the perspective of feature extraction, there are three shortcomings in previous work. First, whether researcher's name appears in URL, title or snippet is a critical factor for homepage identification. It can't be judged by simply string matching. Second, relevance between retrieval results has not been explored, while previous work only considers a single retrieval result. Third, semantic information involved in URL, title or snippet have not been utilized.

3 Academic Homepage Identification

Given a researcher and his affiliation, the query statement will be "researcher name + affiliation", such as "Clark T. C. Nguyen, UC Berkeley Engineering", defined as Q. Each query Q has N retrieval results, named as QR pair.

3.1 Feature Analysis of Academic Homepages

Through the analysis of academic homepages, query statements and retrieval results, we find some typical features and summarize them into four types. These four types of features are described in Table 1.

Query-dependent (QD) Feature. The QR pair's order in the retrieval results is positively related to being a homepage. Thus we extract the QR pair's order as a feature. Other features are from previous work [3], which can be divided into two parts. First, number of researcher's name and researcher's affiliation fragments in URL, title or snippet. Second, keywords related to homepage.

Table 1. Summary of features designed for each QR pair

Type	Function	Feature description	Dim[a]
QD	Order()	QR pair's order in retrieval results of a researcher	1
	Length(U/T/S)	Length of URL/Title/Snippet/RN/RI	5
	Exist(U, special_char)[b]	Each special char(/,=?&-_%∼) exist in URL or not	9
	Exist(U, num_frag)	Pure digital fragments in URL or not	1
	Score(U/T/S, RN_frag)[c]	Score of RN fragments in URL/Title/Snippet	1
	Score(U/T/S, RI_frag)	Score of RI fragments in URL/Title/Snippet	1
	Exist(U, domain)	Each domain name in URL or not	29
	Exist(T/S, keyword)	Each keyword exists in Title/Snippet or not	254
ES	Exist(U/T/S, RN)	Researcher's Name (RN) in URL/Title/Snippet	3
	Exist(U/T/S, RI)	Researcher's Institute (RI) in URL/Title/Snippet	3
LC	Rank(feature_value)	Each feature rank value of a researcher	304
	Norm(feature_value)	Each feature normalized value of a researcher	304
SE	Embed(U/T/S)[d]	Semantic embedding learned from URL/Title/Snippet	256

[a] Dim: Dimension of features.
[b] Exist(s, t): Whether string fragment t exists in s.
[c] Score(s, t): Number of fragments t exists in s divided by the number of fragments.
[d] Embed(seq): Learning semantic embedding from sequence seq.

The keywords contain homepage related topic words such as “students”, “member”, and “committee” extracted from the web-contents of homepages using the Latent Dirichlet Allocation (LDA) model and some high frequency words.

Entity-saliency (ES) Feature. Whether researcher’s name and researcher’s affiliation appear in the retrieval result is quite important for the identification of the homepage. However, researcher’s name has various forms. ES feature aims to make entities more easy to be detected. Hence, some heuristic rules are set to splice name fragments, like from “Jiawei Han” to “JiaweiHan” or from “Charu C. Aggarwal” to “CharuAggarwal”.

Local-contextual (LC) Feature. Existing works only consider a single QR pair, while ignoring the relationship between QR pairs of the same researcher. LC Feature able to establish relations between QR pairs of the same researcher, which contains: Rank: ranks of specific QR pair; Mean: the mean of feature values of top N QR pairs; Variance: the variance of feature values of top N QR pairs; Normalized feature: normalizing feature values by standard deviation.

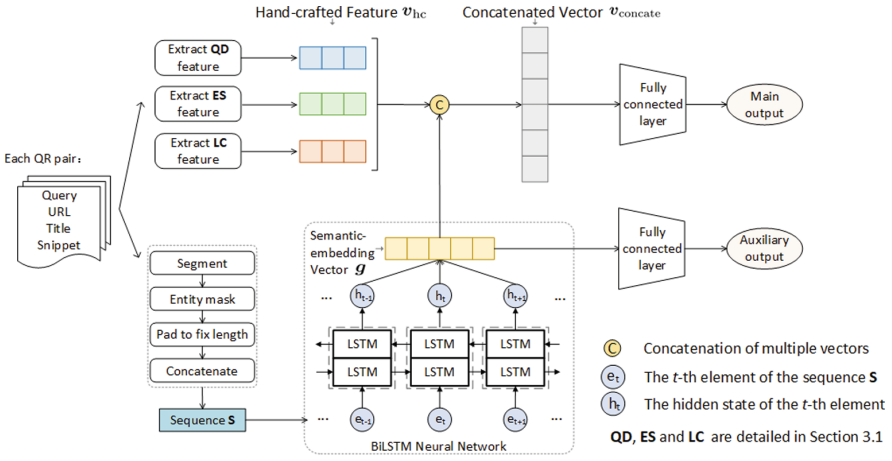

Fig. 1. Architecture of our joint model.

Semantic-embedding (SE) Feature. We first perform segmentation and entity masking operation to URL, title and snippet. Entity masking aims to encode name and institute in a unified way. Then researcher’s name fragment, researcher’s institute fragment and numerical fragment appear could be encoded to three fixed numbers. Sequences of URL, title and snippet are padded to fixed length and concatenated to a single sequence $\boldsymbol{S}$. They were fed into a Bi-directional LSTMs (BiLSTM) [5] model to obtain the semantic information.

3.2 Joint Model in Our Work

In order to integrate hand-crafted features and semantic-embedding feature, we employ a joint neural network and adopt a joint training mode. The joint neural network model, as shown in Fig. 1, aims to improve the identification ability of the model by combining hand-crafted features and semantic information.

The joint neural network model contains two inputs: the first part takes hand-crafted features as input, denoted by $\boldsymbol{v}_{\mathtt{hc}}$; the second part takes sequence information as input, denoted by $\boldsymbol{v}_{\mathtt{seq}}$. Let $\boldsymbol{v}_{\mathtt{seq}}$ input to one layer BiLSTM, the output of BiLSTM layer is a global sentence-level hidden vector $\boldsymbol{g}$ which detailed in Sect. 3.1. The joint neural network model contains two outputs, namely main output and auxiliary output. At the training stage, these two outputs share the same label. At the testing stage, they will output a probability value range from 0 to 1 represent the score of QR pair. The value of main output is regarded as the final score of QR pair.

The joint training model has two advantages. One is to make the BiLSTM and embedding layer being trained smoothly, even if the joint loss value is very high. The other is to make use of semantic information involved in sequences. For a single QR pair i, the loss value can be calculated by formula Eq. (1) and the batch random gradient descent as Eq. (2), m is the batch size. We set a tunable parameter λ to control the joint loss function Eq. (3). The goal of parameter estimation is to find the optimal θ^* to minimize the joint loss function $\mathcal{L}_{\text{joint}}$, y represent the label of the data, $h_\theta(x)$ represent a series of linear or nonlinear transformations.

$$\mathcal{L}(h_\theta(x_i), y_i) = -y_i log(h_\theta(x)) - (1 - y_i) log(1 - h_\theta(x)) \tag{1}$$

$$\mathcal{L}(x, y) = \sum_{i=1}^{m} \mathcal{L}(h_\theta(x_i), y_i) \tag{2}$$

$$\mathcal{L}_{\text{joint}}(x, y) = \lambda \mathcal{L}_{\text{seq}}(x, y) + (1 - \lambda)\mathcal{L}_{\text{hc}}(x, y) \tag{3}$$

4 Experiments

4.1 Experimental Setup

We conduct our experiments on a real-world dataset provided by AMiner[1]. It contains 20,445 researchers and 203,019 corresponding retrieval results where each researcher has 8 to 11 retrieval results. The dataset falls into three parts including a training set (6000 researchers, 59675 QR pairs, 5677 homepages), a validation set (2435 researchers, 24187 QR pairs, 2267 homepages) and a test set (12010 researchers, 119157 QR pairs, 11364 homepages). In order to present impact of our approaches, we set the following 6 groups comparison experiments.

(1) Baseline (BL_SVM/BL_RSVM). This experiment uses features from previous work [3,10], as described in the query-dependent feature part of Sect. 3.1.

[1] https://biendata.com/competition/scholar/.

The SVM [2] and RankSVM [6] models are utilized. (2) Hand-crafted features with SVM (HF_SVM). The SVM model which uses the hand-crafted features proposed in this paper as input. (3) Hand-crafted features with deep neural network (HF_DNN). A deep neural network model which only uses the hand-crafted features to classify the homepages. It could be seen as the baseline of the Joint Neural Network. (4) Semantic-embedding feature with biLSTM (SF_BiLSTM). It's an auxiliary classifier trained to identify the homepages. (5) Combined features with joint neural network (CF_JNN). This experiment combines hand-crafted features and semantic-embedding feature together and adopts joint training mode, detailed in Sect. 3.2.

To better evaluate these approaches, we set three types of evaluation criterion. Accuracy (only if the identified page equals to the labeled homepage, the page is considered to be correct.) precision recall curves [4] and the mean reciprocal rank [9]. For the network configuration, the parameters of the dropout probability are tuned to 0.25, the layer of deep neural network is set to 3, the semantic embedding vector size is set to 256, the batch size is 300 and the optimizer is Adam [7] with a learning rate of 0.001. We implement neural models based on Keras[2] and directly use its default parameter initialization strategy. Since academy homepages are only one tenth of retrieval results, the weight proportion of the positive and negative data is set to 9:1 heuristically to overcome the problem of data imbalance.

Table 2. Accuracies (%) of different approaches.

Method	BL_SVM	BL_RSVM	HF_SVM	HF_DNN	SF_BiLSTM	CF_JNN
Accuracy	59.03	60.17	61.91	67.02	62.96	**71.04**

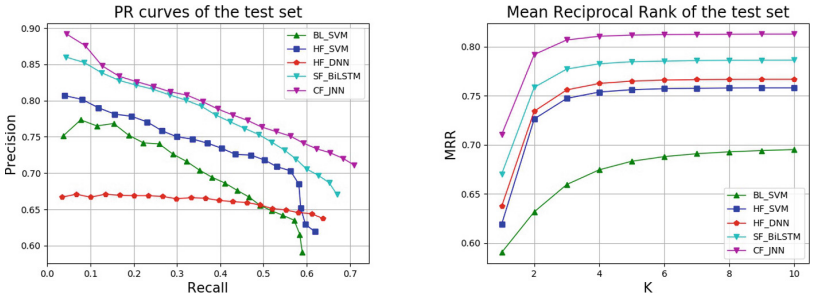

Fig. 2. P-R curves and MRR values of different approaches on test set.

[2] https://keras.io/.

4.2 Experimental Results and Discussion

We first report the accuracy of 6 groups of different experiments in Table 2. Our CF_JNN approach achieves best results. The accuracy is 71.12% in the validation set and 71.04% in the test set. In comparison with baseline approach BL_SVM and BL_RSVM, which accuracy are 60.31% and 60.17%, our approach performs 10–11% better than the baseline. There are two main reasons: one is that our more effective features, which could be proved from the comparison between BL_SVM and HF_SVM; the other is that our joint model and it could be seen from the comparison between HF_DNN and CF_JNN.

From the accuracy of HF_DNN, SF_BiLSTM and CF_JNN, we observe that the joint neural network significantly outperform both SF_BiLSTM and HF_DNN, which means the joint model is effective. According to PR curves in Fig. 2, our improved features and joint neural network are more stable than previous work. From the PR curves of BL_SVM and HF_SVM, although they have similar accuracy, our features have higher F1 values and perform more stable in most cases. The result of the RankSVM model unable to draw the P-R curves, since it's a comparison between retrieval results of the same author instead of giving a global score. These demonstrate the effectiveness of our approach.

5 Conclusion and Further Work

In this paper, we study the problem of academic homepage identification using retrieval results from the search engine. To fully leverage both structural and content information in retrieval results, we propose a joint neural network model to identify academic homepage using both carefully designed features and semantic embeddings. We conduct experiments on a real-world dataset and the experimental results demonstrate the effectiveness of our approach. Our future directions is to investigate the performance of our approach for identifying the related webpages of other entities, such as institute, medicine and weapon.

Acknowledgments. This work was supported in part by the National Key Research and Development Program of China under Grant No. 2016YFB0801003.

References

1. Balog, K., Serdyukov, P., De Vries, A.P.: Overview of the TREC 2010 entity track. Norwegian University of Science and Technology Trondheim, Technical report (2010)
2. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. (TIST) **2**(3), 27 (2011)
3. Das, S., Mitra, P., Giles, C.: Learning to rank homepages for researcher-name queries. In: The International Workshop on Entity-Oriented Search, SIGIR, pp. 53–58. Citeseer (2011)
4. Davis, J., Goadrich, M.: The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 233–240. ACM (2006)

5. Graves, A., Jaitly, N., Mohamed, A.r.: Hybrid speech recognition with deep bidirectional LSTM. In: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 273–278. IEEE (2013)
6. Joachims, T.: Optimizing search engines using clickthrough data. In: ACM Conference on Knowledge Discovery and Data Mining (2002)
7. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
8. Li, H., et al.: CiteSeer χ: a scalable autonomous scientific digital library. In: Proceedings of the 1st International Conference on Scalable Information Systems, p. 18. ACM (2006)
9. Radev, D.R., Qi, H., Wu, H., Fan, W.: Evaluating web-based question answering systems. In: LREC (2002)
10. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: ArnetMiner: extraction and mining of academic social networks. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 990–998. ACM (2008)