Deep Feature Engineering using Full-text Publications

Iqra Safder [1[0000-0001-9818-4693]], Hafsa Batool[1[0000-0002-5664-891X]] and Saeed-Ul Hassan[1[0000-0002-6509-9190]]

[1] Information Technology University, Ferozepur Road, Lahore, Pakistan
iqra.safder@itu.edu.pk

1 Introduction

We have observed a rapid proliferation in scientific literature and advancements in web technologies has shifted information dissemination to digital libraries [1]. In general, the research conducted by scientific community is articulated through scholarly publications pertaining high quality algorithms along other algorithmic specific metadata such as achieved results, deployed datasets and runtime complexity. According to estimation, approximately 900 algorithms are published in top core conferences during the years 2005-2009 [2]. With this significant increase in algorithms reported in these conferences, more efficient search systems with advance searching capabilities must be designed to search for an algorithm and its supported metadata such as evaluation results like precision, recall etc., particular dataset on which an algorithm executed or the time complexity achieved by that algorithm from full body text of an article. Such advanced search systems could support researchers and software engineers looking for cutting edge algorithmic solutions. Recently, state of the art search techniques has been designed to search for an algorithm from full text articles [3-5].

In this work, we designed an advanced search engine for full text publications that leverages the deep learning techniques to classify algorithmic specific metadata and further to improve searching capabilities for a search system.

2 Data, Methodology and Results

We use two data corpuses, a) DS1: a repository of 256 papers downloaded from CiteSeerX containing 275 unique algorithms with 37000 text lines and among them 34,423 lines are marked as negative class samples and 2577 lines as positive class samples. Positive samples are further divided into 3 sub categories and out of total positive instances 2334 lines are tagged as algorithmic efficiency text lines, 136 lines as dataset lines that contains deployed dataset information for an algorithm and remaining 80 lines as algorithmic time complexity lines. DS1 is used to evaluate the performance of our proposed deep learning based technique. b) DS2: a full text repository downloaded from Association of computational linguistics (ACL), containing 21,940 articles with 1500 unique algorithms. We used DS2 as a case study to elaborate the effectiveness of an advanced searching model.

Firstly, pdf full-text articles are converted into plan text and fed to document segmentation step. The purpose of this step is to keep only the relevant sections such as Abstract, Methodology, Experimentation & Results and Conclusion, that may contain discussions related to algorithmic specific metadata and to clip all irrelevant sections (Introduction, Literature Review, Acknowledgement and References). Moreover, we also applied pre-processing steps to remove junk lines, header footers, and author-name and document titles. Next, the cleaned text is fed to our Bi-Directional LSTM based model for algorithmic metadata classification purpose. Lastly, we designed a case study on ACL corpus by generating document synopsis. We enriched these synopses with algorithmic related metadata text lines identified by our designed deep learning model. In the end, we performed an empirical evaluation to show the effectives of Enhanced IR system with Traditional TF-IDF system. For Bi-Directional LSTM, precision and recall results are 0.787 and 0.778 respectively. Overall, our Bi-Directional LSTM-based approach performed well from both the rule-based and machine learning-based approaches, with 79% accuracy.

3 Conclusions

This paper presents an advanced technique for algorithmic metadata extraction in full text scientific publications to enhance searching mechanism in digital libraries.

As a future work, we plan to deploy Machine learning and NLP techniques for contextual and semantic interpretation of numeric text. This kind of work would help to compare algorithms semantically on the basis of their reported evaluation results. Moreover, we would design advanced techniques to effectively extract other useful metadata such as input, output, and compatible data structures of an algorithm and reported code snippet from large scale full text data corpus for improved documents search. The data and code used in this paper can be downloaded from here: https://github.com/slab-itu/ir_icadl_2018.

References

[1] Tuarob, S., Bhatia, S., Mitra, P. and Giles, C.L.: AlgorithmSeer: A system for extracting and searching for algorithms in scholarly big data. IEEE Transactions on Big Data, 2(1), 3-17 (2016).

[2] Bhatia, S. and Mitra, P.: Summarizing figures, tables, and algorithms in scientific publications to augment search results. *ACM Transactions on Information Systems (TOIS)*, *30*(1), 3 (2012).

[3] Safder, I., Hassan, S.U. and Aljohani, N.R.: AI Cognition in Searching for Relevant Knowledge from Scholarly Big Data, Using a Multi-layer Perceptron and Recurrent Convolutional Neural Network Model. In: *Companion of the The Web Conference 2018 on The Web Conference 2018*, pp. 251-258. International World Wide Web Conferences Steering Committee, (2018).

[4] Safder, I., Sarfraz, J., Hassan, S.U., Ali, M. and Tuarob, S.: Detecting Target Text Related to Algorithmic Efficiency in Scholarly Big Data Using Recurrent Convolutional Neural Network Model. In: International Conference on Asian Digital Libraries, pp. 30-40. Springer, Cham (2017).

[5] I. Safder and S.-U. Hassan, "DS4A: Deep Search System for Algorithms from Full-text Scholarly Big Data" In: International Conference on Data Mining Workshop, (2018).