Annotating D3 dataset with the CSO Classifier

Angelo Salatino
angelo.salatino@open.ac.uk
KMi, The Open University
Milton Keynes, UK

Abstract

The DBLP Discovery Dataset (D3) is a newly created dataset of research papers in the field of Computer Science which can support several tasks like identifying trends in research activity, productivity, focus, bias, accessibility, and impact. This dataset stems from DBLP and integrates additional information from the full-texts. We argue that papers classified with their research topics can improve the identification of research trends. To this end, we used the CSO Classifier to annotate all the papers within D3 and we made such extension available for research purposes.

Keywords: Datasets, Annotation, Topic Extraction

1 Introduction

The DBLP Discovery Dataset (D3) is a dataset in the field of Computer Science, which was recently released and can support several tasks including identifying trends in research activity, productivity, focus, bias, accessibility, and impact. This dataset derives from DBLP and integrates additional information from the full-texts [3]. Each paper is associated with a set of attributes: corpusid, abstract, updated, externalids, url, title, authors, venue, year, referencecount, citationcount, influentialcitationcount, isopenaccess, s2fieldsofstudy, publicationtypes, publicationdate, and journal.

We argue that annotating research papers with their research topics can improve a number of tasks, including the exploration of research trends, the recommendation of similar research articles, and extraction of knowledge [2]. To this end, we run the CSO Classifier to annotate all the papers within the D3 dataset and we made such extension available for research purposes on Zenodo[1].

2 CSO Classifier

The CSO Classifier is an application that takes as input the text from abstract, title, and keywords of a research paper and outputs a list of relevant concepts from CSO. It consists of two main components: (i) the syntactic module and (ii) the semantic module. The syntactic module parses the input documents and identifies CSO concepts that are explicitly referred in the document. The semantic module uses part-of-speech tagging to identify promising terms and then exploits word embeddings to infer semantically related topics. Finally, the CSO Classifier combines the results of these two modules, removes outliers, and enhances them by including relevant super-areas. The reader can refer to [1] for additional details.

[1]D3 dataset annotated with CSO topics - https://zenodo.org/record/7097148

3 Dataset

In this section, we will observe how to process the newly created annotation. The D3 dataset is distributed in JSONL format, meaning that each line is a JSON dictionary. This format is quite convenient for large files as it does not require the whole dataset to be parsed at once, but it can be parsed row by row (i.e., paper by paper).

For the sake of consistency, we kept the same format with our annotated dataset.

3.1 D3 dataset

In Listing 1, we present an example of line (paper) found in the D3 dataset, having corpus id 26. In particular, we can observe the richness of metadata pertained in this dataset.

3.2 CSO annotations

In Listing 2 we can find the extracted topics from the same paper (corpus id 26) showed in Listing 1. It is a JSON dictionary that will sit as single line within the distributed dataset. In particular, it contains 5 keys. There is the *corpusid* which helps to refer to the original paper contained in the D3 dataset. Then, there are four keys that express the outcome of the CSO Classifier: *syntactic*, *semantic*, *union*, and *enhanced*. The keys syntactic and semantic respectively contain the topics returned by the syntactic and semantic module. Union contains the unique topics found by the previous two modules. In enhanced you can find the relevant super-areas.

References

[1] Angelo Salatino, Francesco Osborne, and Enrico Motta. 2021. CSO Classifier 3.0: a scalable unsupervised method for classifying documents in terms of research topics. *International Journal on Digital Libraries* (July 2021). https://doi.org/10.1007/s00799-021-00305-y

[2] Angelo A. Salatino, Francesco Osborne, Aliaksandr Birukou, and Enrico Motta. 2019. Improving Editorial Workflow and Metadata Quality at Springer Nature. In *The Semantic Web – ISWC 2019*. Springer International Publishing, Cham, 507–525.

[3] Jan Philip Wahle, Terry Ruas, Saif M. Mohammad, and Bela Gipp. 2022. D3: A Massive Dataset of Scholarly Metadata for Analyzing the State of Computer Science Research. https://doi.org/10.48550/ARXIV.2204.13384

Figure 1. JSON associated to paper (corpusid 26) within the D3 dataset.


```
{
    "corpusid": 26,
    "abstract": "In this paper, we introduce a field-programmable gate array (FPGA) hardware architecture
         for the realization of an algorithm for computing the eigenvalue decomposition (EVD) of para-
        Hermitian polynomial matrices. Specifically, we develop a parallelized version of the second-
        order sequential best rotation (SBR2) algorithm for polynomial matrix EVD (PEVD). The proposed
        algorithm is an extension of the parallel Jacobi method to para-Hermitian polynomial matrices, as
         such it is the first architecture devoted to PEVD. Hardware implementation of the algorithm is
        achieved via a highly pipelined, non-systolic FPGA architecture. The proposed architecture is
        scalable in terms of the size of the input para-Hermitian matrix. We demonstrate the
        decomposition accuracy of the architecture through FPGA-in-the-loop hardware co-simulations.
        Results confirm that the proposed solution gives low execution times while reducing the number of
         resources required from the FPGA.",
    "updated": "2022-02-13T16:00:07.412Z",
    "externalids": {
        "ACL": null,
        "DBLP": "conf/fpt/KasapR12",
        "ArXiv": null,
        "MAG": "1994418445",
        "CorpusId": "26",
        "PubMed": null,
        "DOI": "10.1109/FPT.2012.6412125",
        "PubMedCentral": null
    },
    "url": "https://www.semanticscholar.org/paper/7011b84b03f1d992962c4a6c87459f7742bc3165",
    "title": "FPGA-based design and implementation of an approximate polynomial matrix EVD algorithm",
    "authors": [
        {
            "authorId": "12653318",
            "name": "Server Kasap"
        },
        {
            "authorId": "144237481",
            "name": "Soydan Redif"
        }
    ],
    "venue": "2012 International Conference on Field-Programmable Technology",
    "year": 2012,
    "referencecount": 16,
    "citationcount": 1,
    "influentialcitationcount": 0,
    "isopenaccess": false,
    "s2fieldsofstudy": [
        {
            "category": "Computer Science",
            "source": "s2-fos-model"
        },
        {
            "category": "Computer Science",
            "source": "external"
        }
    ],
    "publicationtypes": [
        "JournalArticle",
        "Conference"
    ],
    "publicationdate": "2012-12-01",
    "journal": {
        "name": "2012 International Conference on Field-Programmable Technology",
        "volume": null,
        "pages": "135-140"
    }
}
```

Figure 2. JSON obtained by the CSO Classifier for the same paper (corpusid 26).


```
{
    "syntactic": [
        "computer hardware",
        "hardware implementations",
        "fpga architectures",
        "proposed architectures",
        "eigenvalue decomposition",
        "field programmable gate array",
        "hardware architecture"
    ],
    "semantic": [
        "field programmable gate array",
        "hardware implementations",
        "programmable gate array",
        "computer hardware",
        "hardware architecture",
        "eigenvalues",
        "eigenvalue decomposition"
    ],
    "union": [
        "computer hardware",
        "hardware implementations",
        "fpga architectures",
        "programmable gate array",
        "proposed architectures",
        "eigenvalue decomposition",
        "field programmable gate array",
        "eigenvalues",
        "hardware architecture"
    ],
    "enhanced": [
        "computer science",
        "logic gates",
        "network architecture",
        "eigenvalues and eigenfunctions",
        "computer networks",
        "matrix algebra",
        "mathematics"
    ],
    "corpusid": 26
}
```