The AIDA Dashboard: Analysing Conferences with Semantic Technologies

Simone Angioni[1], Angelo Salatino[2], Francesco Osborne[2], Diego Reforgiato Recupero[1], and Enrico Motta[2]

[1] Department of Mathematics and Computer Science, University of Cagliari (Italy)
{simone.angioni, diego.reforgiato}@unica.it
[2] Knowledge Media Institute, The Open University, Milton Keynes (UK)
{angelo.salatino, francesco.osborne, enrico.motta}@open.ac.uk

Abstract. Scientific conferences play a crucial role in the field of Computer Science by promoting the cross-pollination of ideas and technologies, fostering new collaborations, shaping scientific communities, and connecting research efforts from academia and industry. However, current systems for analysing research data do not provide a good representation of conferences. Specifically, these solutions do not allow to track research trends, to compare conferences in similar fields, and to analyse the involvement of industrial sectors. In order to address these limitations, we developed the AIDA Dashboard, a tool for exploring and making sense of scientific conferences which integrates statistical analysis, semantic technologies, and visual analytics.

Keywords: Scholarly Data · Knowledge Graphs · Topic Detection · Bibliographic Data · Scholarly Ontologies · Research Dynamics

1 Introduction

Scientific conferences play a crucial role in the field of Computer Science by promoting the cross-pollination of ideas and technologies, fostering new collaborations, shaping scientific communities, and connecting research efforts from academia and industry. For this reason, every significant field is usually associated with multiple conferences that help defining its challenges and paradigms and to coordinate the effort of all the interested stakeholders.

Therefore, understanding and monitoring Computer Science conferences is an important task for editors, researchers, companies, research policy makers and other users working in this space. Several applications and services already provide a wide variety of functionalities to support the exploration of research data and produce various kinds of analytics. These include Microsoft Academic Graph, Semantic Scholar, Scopus, Web of Science, OpenCitations, and many others. However, these systems tend to neglect conferences and offer only a very limited set of relevant analytics, such as the number of papers or citations. In the first instance, they do not allow users to examine the trends of the relevant research topics. It is thus difficult to assess what are the research challenges that a conference is actually addressing and how its focus changed in time. Secondly, there is poor support in comparing conferences to determine the best performing ones in specific fields. For instance, we would like to know which are the main conferences in Semantic Web, how they compare in terms of average citations or other metrics, and how their performance changes in the last few years. A third limitation is that current systems do not report any analytics about the industry involvement. Conversely, it can be argued that conferences are the premium public venues in which industry and academia interact, hence monitoring these dynamics is critical for assessing a conference.

In order to address these limitations, we developed the AIDA Dashboard, a tool for exploring and making sense of scientific conferences which integrates statistical analysis, semantic technologies, and visual analytics. The AIDA Dashboard was developed in collaboration with Springer Nature for assisting editors in assessing conferences, but it also supports several other use cases. It introduces three novel features that state-of-the-art systems are currently lacking. First, it associates to conferences a very granular representation of their topics from the Computer Science Ontology (CSO)[7] [3] and uses it to produce several analytics about its research trends over time. Second, it enables to easily compare and rank conferences according to several metrics within specific fields (e.g., Semantic Web) and time-frames (e.g., last five years). Finally, the AIDA Dashboard offers several features for assessing the involvement of industry in a conference. This includes the ability to focus on companies and their performance when assessing organizations, to report the ratio of publications and citations from academia, industry, collaborative efforts, and to distinguish industrial contributions according to 66 industrial sectors (e.g., automotive, financial, energy, electronics) from the Industrial Sectors Ontology (INDUSO)[4]. A demo of AIDA Dashboard is currently available at `http://w3id.org/aida/dashboard`.

2 The AIDA Dashboard

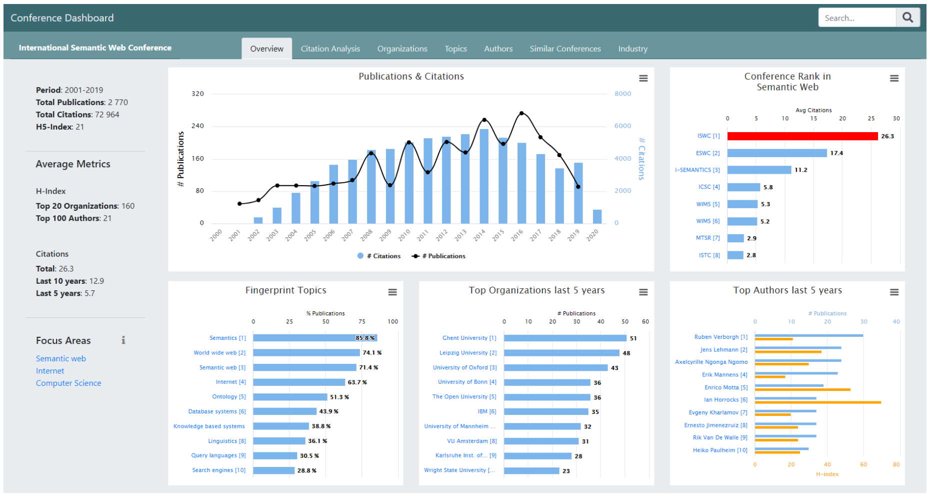
The AIDA Dashboard is a web application that allows users to visualize several kind of analytics about a specific conference (see Figure 1). The backend is developed in Python, while the frontend is in HTML5 and Javascript.

The AIDA Dashboard builds on the Academia/Industry DynAmics [2,1] knowledge graph (AIDA)[5], a large knowledge base describing 14M articles and 8M patents in the field of Computer Science according to the research topics drawn from CSO. 4M articles and 5M patents are also classified according to the type of the author's affiliations (academy, industry, or collaborative) and 66 industrial sectors drawn from INDUSO, which was specifically designed to

[3] CSO - `https://cso.kmi.open.ac.uk/`

[4] INDUSO - `http://w3id.org/aida/downloads/induso.ttl`

[5] AIDA - `http://w3id.org/aida`



Period: 2001-2019
Total Publications: 2 770
Total Citations: 72 964
H5-Index: 21

Average Metrics

H-Index
Top 20 Organizations: 160
Top 100 Authors: 21

Citations
Total: 26.3
Last 10 years: 12.9
Last 5 years: 5.7

Focus Areas
Semantic web
Internet
Computer Science

Fig. 1. The Overview of ISWC according to the AIDA Dashboard.

support AIDA. AIDA was generated by integrating several knowledge graphs and bibliographic corpora, including Microsoft Academic Graph (MAG), Dimensions, DBpedia, CSO, and the Global Research Identifier Database (GRID).

The research papers were annotated with CSO topics using the CSO Classifier [6][6], which is a tool that uses part-of-speech tagging to identify promising terms and then exploits word embeddings to infer semantically related topics from CSO. In addition, to extract further relevant topics, the classifier includes also all their super topics according to the CSO. For instance, a paper tagged with *Neural Networks* would be assigned the topic *Artificial Intelligence*. This solution enables identifying high level topics that are not typically mentioned in the documents. The CSO Classifier powers the current version of the Smart Topic Miner [5], which is the application used by Springer Nature to semi-automatically annotate Proceedings books in the field of Computer Science. Since CSO is often updated, the set of topics used by AIDA is also evolving, constantly including new emerging topics. As an example, topics can be extended by using hyperlinks present in papers that might become Semantic Web entities or properties [3].

Each research article was also linked to the industrial sectors described in INDUSO by mapping the affiliations of the authors to their DBpedia entities, which in turn are mapped to INDUSO. For instance, an article that was written by authors who have Toyota as affiliation would be associated to the industrial sector *Automotive*.

AIDA is available at `http://w3id.org/aida` under the CC-BY 4.0 license. It was recently used for supporting the generation of adavanced analytics about research dynamics and forecasting the impact of research topics on industry [4]. However, using these data was not easy for less technical-savvy users. AIDA

[6] CSO Classifier - `https://pypi.org/project/cso-classifier/`

Dashboard is the first step in allowing users to access AIDA through a user-friendly but comprehensive interface.

In order to support the AIDA Dashboard, we pre-computed a full set of analytics for each conference from AIDA-KG and store it in a JSON file that will be loaded by the web interface. This solution allows AIDA Dashboard to be extremely scalable, since for a given conference it needs to query the server only once, to retrieve its associated file. Every other operation is handled by the front-end.

AIDA Dashboard is highly scalable and allows to browse the different facets of a conference according to seven tabs: *Overview*, *Citation Analysis*, *Organizations*, *Authors*, *Topics*, *Similar Conferences*, and *Industry*.

Figure 1 shows the **Overview** tab. This is the main view of a conference that provides introductory information about its performance, the main authors and organization, and the conference rank in its main fields in terms of average citations for paper during the last five years.

The **Citation Analysis** tab reports the evolution in time of several citation-based metrics such as the impact factor and the average citations for paper. It also shows the evolution of the rank and the percentile of the conference in different fields. For instance, the Conference on Neural Information Processing Systems (NeurIPS) is currently the second conference in terms of average citations in Neural Network, the third in Machine Learning, and the twelfth in Artificial Intelligence. This visualization is typically used by Springer Nature editors to assess the performance of conferences within different communities and to identify emerging conferences.

The **Organizations** and **Authors** tabs show several analytics about the main institutions and researchers active in the conference. Organizations can be filtered according to their type (academia or industry) and are associated with their number of publications, citations, and average citations for paper. The researchers are associated with similar analytics, but also with their H-index and H5-index, in order to quickly identify high impact researchers. Editors use this information to understand the quality of researchers and organizations attracted by the conferences. This is particularly important for assessing relatively young conferences that may not have developed yet a strong citation record.

The **Topic** tab allows users to analyse the topic trends in time. Specifically it shows two selections of topics: frequent topics and fingerprint topics. The first is the set of topics which appear more frequently in the conference. The second is the set of most distinctive topics of the conference. It is obtained by computing the difference between the topic distribution of the conference and the one of the full dataset. Preliminary analyses revealed that this second set is usually able to better represent the topics considered central to the conference.

The **Similar Conferences** tab compares the conference under analysis with all the other conferences in the same fields according to their number of publications, citations, and average citations for paper. The user can contextualise the comparison to different fields. For example, ISWC can be compared with all the other conferences in the fields of Semantic Web, Internet, or Computer Science.

Finally, the **Industry** tab reports the percentage of publications and citations from academia, industry, and collaborative efforts as well as the industrial sectors analysis. The latter shows the percentage of produced publications and citations received by companies in different sectors. For instance, the main industrial sectors of ISWC are *Computing and IT*, *Information Technology*, *Management*, *Telecommunication*, and *Health Care*.

3 Conclusions

The current version of AIDA Dashboard already provides an array of interesting functionalities, many of which go beyond what is available in other current tools. Nevertheless, we are still at a relatively early stage and we are planning to introduce new ones. As first step, we plan to add a geographical tab for analysing the distribution of countries active in a conference. We also want to expand the set of entities that could be analysed by the dashboard, producing similar analytics also for journals, organizations, and scientific communities. Finally, we plan to perform a comprehensive user study with editors and researchers from different communities in order to assess the system and collect useful feedback. For such a purpose, to generalize the presented dashboard we only need to replace the CSO ontology with others within the domain under study. As such, we have already started working with the MeSH ontology within the bio-informatics domain to have our dashboard working in that domain as well.

References

1. Angioni, S., Osborne, F., Salatino, A.A., Recupero, D.R., Motta, E.: Integrating knowledge graphs for comparing the scientific output of academia and industry. In: Proc. of the ISWC 2019 Satellite Tracks. CEUR Workshop Proceedings, vol. 2456, pp. 85–88 (2019)
2. Angioni, S., Salatino, A., Osborne, F., Reforgiato Recupero, D., Motta, E.: Integrating knowledge graphs for analysing academia and industry dynamics. In: ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium. Springer International Publishing, Cham (2020)
3. Presutti, V., Nuzzolese, A.G., Consoli, S., Gangemi, A., Recupero, D.R.: From hyperlinks to semantic web properties using open knowledge extraction. Semantic Web **7**(4), 351–378 (2016). https://doi.org/10.3233/SW-160221
4. Salatino, A., Osborne, F., Motta, E.: Researchflow: Understanding the knowledge flow between academia and industry. In: Knowledge Engineering and Knowledge Management. Springer International Publishing (2020)
5. Salatino, A.A., Osborne, F., Birukou, A., Motta, E.: Improving editorial workflow and metadata quality at springer nature. In: The Semantic Web – ISWC 2019. pp. 507–525. Springer International Publishing, Cham (2019)
6. Salatino, A.A., Osborne, F., Thanapalasingam, T., Motta, E.: The cso classifier: Ontology-driven detection of research topics in scholarly articles. In: Digital Libraries for Open Knowledge. pp. 296–311. Springer International Publishing, Cham (2019)
7. Salatino, A.A., Thanapalasingam, T., Mannocci, A., Osborne, F., Motta, E.: The computer science ontology: a large-scale taxonomy of research areas. In: International Semantic Web Conference. pp. 187–205. Springer (2018)