AIDA: a Knowledge Graph about Research Dynamics in Academia and Industry

Simone Angioni¹, Angelo Salatino², Francesco Osborne², Diego Reforgiato Recupero¹, and Enrico Motta²

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. Academia and industry share a complex, multifaceted, and symbiotic relationship. Analysing the knowledge flow between them, understanding which directions have the biggest potential, and discovering the best strategies to harmonise their efforts is a critical task for several stakeholders. While research publications and patents are an ideal media to analyse this space, current datasets of scholarly data can not be used for such a purpose since they lack a high-quality characterization of the relevant research topics and industrial sectors. In this paper, we introduce the Academia/Industry DynAmics (AIDA) Knowledge Graph, which describes 14M publications and 8M patents according to the research topics drawn from the Computer Science Ontology. 4M publications and 5M patents are further characterized according to the type of the author’s affiliations (academia, industry, or collaborative) and 66 industrial sectors (e.g., automotive, financial, energy, electronics) organized in a two-level taxonomy. AIDA was generated by means of an automatic pipeline that integrates data from Microsoft Academic Graph, Dimensions, DBpedia, the Computer Science Ontology, and the Global Research Identifier Database. It is publicly available under CC BY 4.0 and can be downloaded as a dump or queried via a triplestore. We evaluated the parts of the AIDA pipeline on a manually crafted gold standard yielding competitive results.

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

1 Introduction

Academia and industry share a complex, multifaceted, and symbiotic relationship. Their collaboration and exchange of ideas, resources, and persons is conducive to the production of new knowledge that will ultimately shape the society of the future. Analysing the knowledge flow between academia and industry, understanding which directions have the biggest potential, and discovering the best strategies to harmonise their efforts is thus a critical task for governments, funding bodies, and other stakeholders. For instance, governments and funding
agencies regularly measure the impact of research efforts at several levels and take decision about which ones to fund. Companies need to monitor important research developments, which may offer a competitive advantage. Researchers must keep up with the latest trends and be aware of complementary research efforts from the industrial sector.

The relationship between academia and industry has been analysed from several perspectives in the literature, focusing for instance on the characteristics of direct collaborations [3], the influence of industrial trends on curricula [31], and the quality of the knowledge transfer [3]. However, most of the quantitative studies on this relationship were limited to small-scale datasets or focused on very specific research questions [6,2].

Research publications and patents are an ideal media to analyse the knowledge generated and developed by academia and industry [34]. Today, we have several large-scale knowledge graphs which describe research articles according to their titles, abstracts, authors, organizations, and other metadata. Examples include Microsoft Academic Graph [3], Scopus [4], Semantic Scholar [5], Aminer [6], Core [7], OpenCitations [8], and others. Other resources, such as Dimensions [9], the United States Patent and Trademark Office (USPTO) [10], the Espacenet dataset [11], and the PatentScope corpus [12], offer a similar description of patents. However, these datasets cannot be directly used to analyse the research dynamics of academia and industry since they lack a high quality characterization of the relevant research topics and industrial sectors. First, current solutions do not allow us to easily discriminate if a document (research paper or patent) is from academia or industry. Second, they typically offer a coarse-grained characterization of research topics, which are usually represented only as a flat list of keywords chosen by the authors or extracted from the abstract. This purely syntactic solution is unsatisfactory [18], as it fails: i) to distinguish research topics from other generic keywords; ii) to deal with situations where multiple labels exist for the same research area; and iii) to model and take advantage of the semantic relationships that hold between research areas. For instance, we want to be able to infer that all documents tagged with the topic Neural Network are also about Machine Learning and Artificial Intelligence. This richer representation would allow us to retrieve all the publications which address the concept “artificial intelligence”. Another issue is that current scholarly datasets do not characterize companies according to their sectors. Therefore, it is not possible to measure the impact of a topic (e.g., sentiment analysis, deep learning, semantic web) on different kinds of industry (e.g., automotive, financial, energy).

---

3 Microsoft Academic Graph - [http://aka.ms/microsoft-academic](http://aka.ms/microsoft-academic)
4 Scopus - [https://www.scopus.com/](https://www.scopus.com/)
5 Semantic Scholar - [https://www.semanticscholar.org/](https://www.semanticscholar.org/)
6 Dimensions - [https://www.dimensions.ai/](https://www.dimensions.ai/)
7 USPTO - [https://www.uspto.gov/](https://www.uspto.gov/)
8 Espacenet dataset - [https://worldwide.espacenet.com/](https://worldwide.espacenet.com/)
9 PatentScope - [https://patentscope.wipo.int/](https://patentscope.wipo.int/)
In this paper, we introduce the Academia/Industry DynAmics (AIDA) Knowledge Graph, which describes 14M publications and 8M patents according to the research topics drawn from the Computer Science Ontology. 4M publications and 5M patents are further characterized according to the type of the author’s affiliations (academia, industry, or collaborative) and 66 industrial sectors (e.g., automotive, financial, energy, electronics) organized in a two level taxonomy.

AIDA was generated by means of an automatic pipeline that integrates data from Microsoft Academic Graph (MAG), Dimensions, English DBpedia, the Computer Science Ontology (CSO), and the Global Research Identifier Database (GRID), respectively containing information about 208M research papers, 38M patents, 4.58M entities, 14K research topics, and 97K organisations. In order to represent the industrial sectors of these documents, we also designed the Industrial Sectors Ontology (INDUSO), which describes 66 high-level industrial sectors organized on two levels. We evaluated the different parts of the pipeline for generating AIDA on a manually crafted gold standard yielding competitive results.

The resulting knowledge base enables analysing the evolution of research topics across academia and industry and study the characteristics of several industrial sectors. For instance, it enables detecting the research trends most interesting for the automotive sector or which prevalent industrial topics were recently adopted and investigated by academia. This information can also be used to train forecaster systems for predicting the impact of topics or technologies [17]. Finally, AIDA can be used to characterize authors, citations, countries, and several other entities in MAG according to their topics and industrial sectors. This makes possible to study further dynamics such as the migration of researchers and the citation flow between academia and the industry.

AIDA is available at http://aida.kmi.open.ac.uk. It can be downloaded as a dump or queried via a Virtuoso triplestore at http://aida.kmi.open.ac.uk/sparql/. We plan to release a new version of AIDA every six months, in order to regularly update the publications, the topics, and the company types.

The rest of the paper is organised as follows. In Section 2, we review the literature on current datasets for studying and quantifying the relationship between academia and industry. In Section 3, we describe the pipeline to generate AIDA, give an overview of the resulting knowledge graph, and discuss our strategy for releasing new versions. Section 4 reports the evaluation of the different parts of the AIDA pipeline. Finally, in Section 5 we summarise the main conclusions and outline future directions of research.

2 Literature Review

In the last years we saw the emergence of a quantity of knowledge graphs describing research publications and their metadata. Microsoft Academic Graph (MAG) [30] is a heterogeneous knowledge graph containing the metadata of more than 300M scientific publications, including citations, authors, institutions, journals, conferences, and fields of study. The Semantic Scholar Open Research
Corpus[10] is a dataset of about 185M publications released by Semantic Scholar, an academic search engine provided by the Allen Institute for Artificial Intelligence (AI2). The OpenCitations Corpus [21] is released by OpenCitations, an independent infrastructure organization for open scholarship dedicated to the publication of open bibliographic and citation data with semantic technologies. The current version includes 55M publications and 655M citations. Scopus is a well-known dataset curated by Elsevier, which includes about 70M publications and is often used by governments and funding bodies to compute performance metrics. The AMiner Graph [33] is the corpus of more than 200M publications generated and used by the AMiner system[11] AMiner is a free online academic search and mining system which also extract researchers’ profiles from the Web and integrates them in the metadata. The Open Academic Graph (OAG)[12] is a large knowledge graph integrating Microsoft Academic Graph and AMiner Graph. The current version cover contains 208M papers from MAG and 172M from AMiner. Core[13] is a repository that integrates 24M open access research outputs from repositories and journals worldwide. The Dimensions Corpus is a dataset produced by Digital Science which integrates and interlink 109M research publications, 5.3M grants, and 40M patents. Publications and citations are freely available for personal, non-commercial use.

We decided to adopt MAG over the alternatives for two main reasons. First, it appears to be the most comprehensive among the publicly available datasets of publications. Second, it associates articles with DOIs and organizations with GRID identifiers and therefore can be easily integrated with other knowledge graphs.

Several other resources focus specifically on patents [27]. For instance, the European Patent Office (EPO) curates the Espacenet dataset, which currently covers about 110 million patents from all over the world. Similarly, the United States Patent and Trademark Office produces a corpus which includes more than 14M US patents. The World Intellectual Property Organization (WIPO) offers instead the PatentScope dataset, which contains 84M patent documents, including 4M international patent applications.

We choose Dimensions for AIDA because of its comprehensiveness and also it identifies organizations with GRID IDs, allowing us to easily integrate them with MAG affiliations.

Another category of scientific knowledge graphs comprises those that include a semantic representation of the content of scientific articles. The Semantic Web community has been working for a while on this direction, fostering the Semantic Publishing paradigm [28], creating bibliographic repositories in the Linked Data Cloud [16], generating knowledge bases of biological data [5], formalising research workflows [32], extracting knowledge graphs from research papers [18,8], implementing systems for managing nano-publications [11,14] and micropublica-

\[\text{ORC} \rightarrow \text{https://s2-public-api-prod.us-west-2.elasticbeanstalk.com/corpus/}\]
\[\text{AMiner} \rightarrow \text{https://www.aminer.cn/}\]
\[\text{Open Academic Graph} \rightarrow \text{https://www.openacademic.ai/oag/}\]
\[\text{Core} \rightarrow \text{https://core.ac.uk/}\]
tions, and developing a variety of ontologies to describe scholarly data, e.g., SWRC, BIBO, BiDO, FABIO, SPAR, CSO, and SKGO. Several of these knowledge bases focus on describing the research areas of scientific publications. These include the Medical Subject Heading (MeSH in Biology, Mathematics Subject Classification (MSC) in Mathematics, Physics Subject Headings (PhySH) in Physics, and many others.

In the field of Computer Science, the best-known taxonomies of research areas are the ACM Computing Classification System and the Computer Science Ontology (CSO). The first one is developed and maintained by the Association for Computing Machinery (ACM). It contains around 2K concepts and it is manually curated. Conversely, CSO is automatically generated from a large collection of publications by the Open University and includes about 14K research areas. We adopted CSO for AIDA because it is one order of magnitude larger than the alternatives and it comes with the CSO Classifier, which is a tool for automatically annotating documents with CSO topics. It thus allows us to easily generate a granular representation of all the documents integrated from MAG and Dimensions.

3 AIDA: Academia Industry DynAmics Knowledge Graph

The Academia/Industry DynAmics (AIDA) Knowledge Graph includes 652M triples which describe a large collection of publications and patents in Computer Science according to the author’s affiliations (academia, industry, or collaborative), research topics, and industrial sectors. Specifically, it describes 14M publications from MAG and 8M patents from Dimensions according to the research topics drawn from the Computer Science Ontology (CSO). On average, each publication is associated with 27 ± 19 topics and each patent with 33 ± 14. The 4M publications and 5M patents that were associated with GRID IDs in the original data, are also classified according to the type of the author’s affiliations (academia, industry, or collaborative) and 66 industrial sectors (e.g., automotive, financial, energy, electronics) drawn from the Industrial Sectors ontology (INDUSO) which was specifically designed to support AIDA. AIDA was gen-

---

14 SWRC - http://ontoware.org/swrc
15 BIBO - http://bibliontology.com
16 BiDO - http://purl.org/spar/bido
17 FABIO - http://purl.org/spar/fabio
18 SPAR - http://www.sparontologies.net/
19 CSO - https://cso.kmi.open.ac.uk/
20 SKGO - https://github.com/saidfathalla/Science-knowledge-graph-ontologies
21 Medical Subject Heading - https://www.ncbi.nlm.nih.gov/mesh
22 Mathematics Subject Classification - https://mathscinet.ams.org/msc
23 Physics Subject Headings - https://physh.aps.org/
24 ACM Classification System - https://www.acm.org/publications/class-2012
erated and it will be regularly updated by an automatic pipeline that integrates
and enriches data from Microsoft Academic Graph (MAG), Dimensions, English
DBpedia, the Global Research Identifier Database (GRID), CSO, and INDUSO.

<table>
<thead>
<tr>
<th>Table 1. AIDA - Number of documents in different categories.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Patents</td>
</tr>
</tbody>
</table>

Table 1 shows the number of publications and patents from academia, in-
dustry, and collaborative efforts. When considering the typed documents, most
publications (78.4%) are written by academic institutions, however industry con-
tributes to a good number of them (18.8%). The situation is reversed when con-
sidering patents: 97% of them are from industry and only 2.7% from academia.
Another interesting finding is that the collaborative efforts are limited, counting
only 2.8% of the publications and 0.3% of the patents. These numbers require
further analysis but may suggest that we need to improve the mechanisms to
support and fund collaborative works.

The data model of AIDA builds on SKOS, CSO, and INDUSO and it is
available at [http://aida.kmi.open.ac.uk/ontology](http://aida.kmi.open.ac.uk/ontology). It focuses on four types
of entities: publications, patents, topics, and industrial sectors. In order to en-
able full compatibility with other knowledge graphs (e.g., MAG, Scopus, DBLP,
Semantic Scholar), publications are identified according to their Digital Object
Identifier (DOI) and patents according to their World Intellectual Property Or-
ganization (WIPO) ID. Publications are also associated with authors (identified
with MAG IDs) and organizations (identified with GRID IDs).

The main information about publications and patents are given by mean of
the following semantic relations:

- `hasTopic`, which associates to the documents all their relevant topics drawn
  from CSO;
- `hasAffiliationType`, `hasOriginalAssigneeType`, and `hasCurrentAssigneeType`,
  which associate to the documents the three categories (academia, industry,
  or collaborative) describing the affiliations of their authors (for publications)
  or of the original or current assignees (for patents);
- `hasIndustrialSector`, which associates to documents and affiliations the rele-
  vant industrial sectors drawn from INDUSO.

AIDA includes also some additional relationships which support more specific
analyses:

- `hasSyntacticTopic` and `hasSemanticTopic`, which give respectively all the
  topics extracted using the syntactic module and the semantic module of
  the CSO Classifier [22]. The first set is composed by topics that are explic-
  itly mentioned in the documents. It thus have high precision but low recall,
and may be used by applications for which precision is paramount. The second one consists in topics that do not directly appear in the text, but were inferred using word embeddings.

- hasPercentageOfAcademia and hasPercentageOfIndustry, which respectively give the percentage of authors from academia and industry. It may be used to generate analytics that need to further segment the collaborative category.
- schema:creator and schema:memberOf\(^{26}\) that link documents to authors and authors to affiliations.

- hasGridType which associates the eight categories of organizations described in GRID (Education, Healthcare, Company, Archive, Nonprofit, Government, Facility, and Other) to documents and affiliations.

AIDA is accessible via a Virtuoso triplestore at \(\text{http://aida.kmi.open.ac.uk/sparql}\). A dump of AIDA is also available at \(\text{http://aida.kmi.open.ac.uk/}\). AIDA is is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0) meaning that everyone is allowed to: i) copy and redistribute the material in any medium or format; ii) remix, transform, and build upon the material for any purpose, even commercially.

In the following subsections, we will discuss the automatic generation of AIDA (Section 3.1), present a brief overview (Section 3.2), and describe our plan for producing new versions (Section 3.3).

### 3.1 AIDA Generation

The automatic pipeline for generating AIDA works in three steps: topics detection, extraction of affiliation types, and industrial sector classification. In the following we will detail each phase of the process.

**Topic Detection.** We first collect all the publications and patents from MAG and Dimensions within the Computer Science domain. In particular, we extract the papers from MAG classified as “Computer Science” in their Field of Science (FoS)\(^{29}\), an in-house taxonomy of research domains developed by Microsoft. Similarly, the patents in Dimensions are classified according to the International Patent Classification (IPC) and the fields of research (FoR) taxonomy, which is part of the Australian and New Zealand Standard Research Classification (ANZSRC). To extract only the patents from the Computer Science domain, we select those with the following IPC classification: “Computing, Calculating or Counting” (G06), “Educating, Cryptography, Display, Advertising, Seals” (G09), “Information Storage” (G11), “Information and Communication Technology” (G16), and others (G99). We also select those having the following field of research: “Information and Computing Science” (08), and “Technology” (10). The resulting dataset includes 14M publications and 8M patents.

Since the fields of study in MAG and fields of research in Dimensions are not specific enough for a detailed analysis of the knowledge flow, we then annotate

\(^{26}\) Respectively \(\text{https://schema.org/creator}\) and \(\text{https://schema.org/memberOf}\)
each document with the research topics from CSO. In order to do this, we run the CSO Classifier\textsuperscript{27} on the title and the abstract of all research papers and patents and augment the resulting set of research topics by also including all their super-topics according to the superTopicOf\textsuperscript{28} relationship in CSO. The CSO Classifier is an unsupervised method which exploits syntactic and semantic strategies to classify a document with a set of topics from CSO.

**Extraction of Affiliation Types.** In the second step, we classify papers and patents according to the nature of the relevant organizations in the GRID database. Both MAG and Dimensions link organizations to their GRID ID. In turn, GRID associates each ID with geographical location, date of establishment, alternative labels, external links, and type of institution (e.g. Education, Healthcare, Company, Archive, Nonprofit, Government, Facility, Other). We leverage this last field to tag each document as ‘academia’, ‘industry’, or ‘collaborative’. A document is assigned an ‘academia’ type if all the authors or original assignees have an educational affiliation (university, research center), an ‘industry’ type if they have an industrial affiliation, and a ‘collaborative’ type if there is at least one creator from academia and one from industry. AIDA includes also the other categories from GRID through the relation hasGridType.

**Industrial Sector Classification.** In order to characterize the industrial sectors addressed by each document we designed the Industrial Sector Ontology (INDUSO), which is a two-level taxonomy describing 66 sectors and their relationships. INDUSO was created using a bottom-up method that took in consideration the large collection of publications and patents from MAG and Dimensions. Specifically, for each affiliation described in the documents with a GRID ID, we extracted from DBpedia the objects of the properties “About:Purpose” and “About:Industry”. This resulted in a noisy and redundant sets of 699 sectors. We then clustered together similar industrial sectors. For instance, the industrial sector “Computing and IT” was derived by categories such as “Networking hardware”, “Cloud Computing”, and “IT service management”. Finally, we arranged the resulting sectors in a two level taxonomy with the purpose of having the first level sector including different second level sectors. For example, the first level sector “energy” includes “nuclear power”, “oil and gas industry”, and “air conditioning”. Specifically, the INDUSO ontology contains the following properties:

- the skos:broader property links the first level sectors to the second level sectors.
- the prov:wasDerivedFrom property associated each of the 66 industrial sector to the original 699 sectors they were derived from DBpedia.
- the rdf:type property is used to define the 66 sectors as :industrialSector and the original 699 sectors as :DBpediaCategory.

\textsuperscript{27} CSO Classifier - https://pypi.org/project/cso-classifier/
\textsuperscript{28} CSO Schema - https://cso.kmi.open.ac.uk/schema/cso
In order to tag documents with INDUSO, we associate to it all the industrial sectors that were derived from the BDpedia representation of its organizations. For instance, the documents with an author affiliation described in DBpedia as ‘natural gas utility’ are tagged with the second level sector ‘Oil and Gas Industry’ and the first level sector ‘Energy’.

### 3.2 AIDA Overview

In this section we present an overview of AIDA and discuss some exemplary analytics supported by this resource.

Figure 1 shows the most frequent 16 high level topics (direct sub-topics of Computer Science in CSO) and reports the relevant percentage of academic publications, industry publications, academic patents, and industry patents. Some topics, such as Artificial Intelligence and Theoretical Computer Science, are most addressed by academic publications. Other ones, e.g., Computer Security, Computer Hardware, and Information Retrieval attract a stronger interest from industry. The topics which are mostly associated with patents are Computer Networks, Internet, and Computer Hardware.

![Fig. 1. Distribution of the main topics.](image)

Figure 2 shows the percentage of publications from academia (A) and industry (I) for the same 16 topics across three windows of time (1989-1998, 1999-2008, and 2009-2018). Some evident trends include the sharp growth of Computer Security, Information Retrieval, Computer Network, and Internet. Some other topics, such as Software Engineering and Computer Aided Design appear to become less prolific over the last years. We refer the reader interested in a more granular analysis of these topic trends to Mannocci at al. [15] that used a preliminary version of this dataset for detecting several emerging topics in the field of Human-Computer Interaction.

Figure 3 shows the distribution of the most popular 5K topics according to the academia-industry (horizontal axis) and papers-patents (vertical axis) indexes. The papers-patents index of a certain topic $t$ is the difference between the number of research papers and patents related to $t$, over the whole set of
documents. If this index is positive a topic tends to be associated with a higher number of publications, while if it is negative with a higher number of patents. On the other hand, the academia-industry index for a certain topic $t$ is the difference between the documents in academia and industry related to $t$, over the whole set of documents. If this index is positive a topic tends to be mostly associated with academia, if it is negative with industry.

The topics are tightly distributed around the bisector: the ones which attract most interest from academia are prevalently associated with publications (top-right quadrant), while the ones in industry are mostly associated with patents (bottom left quadrant). These results are consistent with previous analyses in the field [10].

Figure 4 shows the percentage of publications and patents associated with the most prominent industrial sectors. Since AIDA mainly covers Computer Science, the most popular sectors (e.g., Technology, Computing and IT, Electronics, and
Telecommunications, and Semiconductors) are linked to this field. However, we can also appreciate the solid presence of sectors such as Financial, Health Care, Transportation, Home Appliance, and Editorial.

Table 2. Topic composition of some prominent industrial sectors. In bold the highest value for each row.

<table>
<thead>
<tr>
<th></th>
<th>Computing and IT</th>
<th>Telecommunications</th>
<th>Electronics</th>
<th>Semiconductor</th>
<th>Inf. Technology</th>
<th>Photography</th>
<th>Automotive</th>
<th>Financial</th>
</tr>
</thead>
<tbody>
<tr>
<td>Artificial Intelligence</td>
<td>-22%</td>
<td>-27%</td>
<td>-3%</td>
<td>3%</td>
<td>-13%</td>
<td>0%</td>
<td>-23%</td>
<td>-8%</td>
</tr>
<tr>
<td>Computer Aided Design</td>
<td>-22%</td>
<td>-27%</td>
<td>-3%</td>
<td>3%</td>
<td>-13%</td>
<td>0%</td>
<td>-23%</td>
<td>-8%</td>
</tr>
<tr>
<td>Computer Hardware</td>
<td>-7%</td>
<td>-7%</td>
<td>-7%</td>
<td>11%</td>
<td>-5%</td>
<td>0%</td>
<td>-5%</td>
<td>-5%</td>
</tr>
<tr>
<td>Computer Network</td>
<td>-3%</td>
<td>-9%</td>
<td>-9%</td>
<td>11%</td>
<td>-9%</td>
<td>-4%</td>
<td>-15%</td>
<td>-8%</td>
</tr>
<tr>
<td>Computer Programming</td>
<td>-8%</td>
<td>-29%</td>
<td>-1%</td>
<td>32%</td>
<td>-52%</td>
<td>31%</td>
<td>-16%</td>
<td>-32%</td>
</tr>
<tr>
<td>Computer Security</td>
<td>6%</td>
<td>-1%</td>
<td>1%</td>
<td>-27%</td>
<td>-1%</td>
<td>9%</td>
<td>-15%</td>
<td>21%</td>
</tr>
<tr>
<td>Computer System</td>
<td>1%</td>
<td>-1%</td>
<td>1%</td>
<td>0%</td>
<td>4%</td>
<td>-2%</td>
<td>-17%</td>
<td>-10%</td>
</tr>
<tr>
<td>Computer Vision</td>
<td>7%</td>
<td>-1%</td>
<td>21%</td>
<td>-16%</td>
<td>-20%</td>
<td>44%</td>
<td>-7%</td>
<td>54%</td>
</tr>
<tr>
<td>Data Mining</td>
<td>28%</td>
<td>-29%</td>
<td>17%</td>
<td>-35%</td>
<td>49%</td>
<td>-18%</td>
<td>-34%</td>
<td>-17%</td>
</tr>
<tr>
<td>Human-computer interaction</td>
<td>14%</td>
<td>-9%</td>
<td>4%</td>
<td>-41%</td>
<td>9%</td>
<td>-21%</td>
<td>-9%</td>
<td>32%</td>
</tr>
<tr>
<td>Information Retrieval</td>
<td>6%</td>
<td>-16%</td>
<td>14%</td>
<td>-55%</td>
<td>-5%</td>
<td>73%</td>
<td>-37%</td>
<td>20%</td>
</tr>
<tr>
<td>Information Technology</td>
<td>29%</td>
<td>-13%</td>
<td>4%</td>
<td>-51%</td>
<td>50%</td>
<td>33%</td>
<td>41%</td>
<td>-9%</td>
</tr>
<tr>
<td>Internet</td>
<td>3%</td>
<td>-4%</td>
<td>0%</td>
<td>2%</td>
<td>-1%</td>
<td>1%</td>
<td>-9%</td>
<td>-24%</td>
</tr>
<tr>
<td>Operating System</td>
<td>14%</td>
<td>-8%</td>
<td>3%</td>
<td>1%</td>
<td>61%</td>
<td>-24%</td>
<td>-55%</td>
<td>-9%</td>
</tr>
<tr>
<td>Robotics</td>
<td>3%</td>
<td>-1%</td>
<td>10%</td>
<td>-14%</td>
<td>-9%</td>
<td>-18%</td>
<td>21%</td>
<td>15%</td>
</tr>
<tr>
<td>Software Engineering</td>
<td>-22%</td>
<td>-16%</td>
<td>4%</td>
<td>2%</td>
<td>55%</td>
<td>-24%</td>
<td>-20%</td>
<td>-31%</td>
</tr>
</tbody>
</table>

AIDA also enables to analyze how these sectors have a different composition in regards to research topics. Table 2 highlights the key topics of a set of exemplary sectors by reporting the difference between the normalised number of publications in a sector and overall. The darker cells mark the main topics for each sector. For instance, the publications written by authors from the Semiconductor sector refer to the topics Computer Aided Design 90% more frequently than the average publication.

The industrial sectors have a very distinct composition, even when considering just the high-level topics in the table. For instance the Automotive sector focus mainly on Robotics, Software Engineering, and Artificial Intelligence; the Telecommunications sector mainly focuses on Computer Network, Internet, and Computer Hardware; and the Photography sector on Information Retrieval, Computer Vision, and Artificial Intelligence.
3.3 Generation of new updates

We plan to periodically release new versions of AIDA, which will include the most recent publications and patents, as well as the latest versions of CSO and INDUSO. Specifically, we will run the pipeline described in this section over a new dump of documents from MAG and Dimensions every six months. In addition, we also plan to release a new version whenever a significant new version of CSO or INDUSO is produced.

4 Evaluation of AIDA

The following sub-sections describe the evaluations performed for assessing the topic classification, the academia/industry classification, and the industrial sector classification.

4.1 Topic Classification

We compared the CSO Classifier, which we use to annotate documents according to their topics, against thirteen approaches using a gold standard made of 70 most cited papers [22] within the fields of Natural Language Processing (23 papers), Semantic Web (23), and Data Mining (24). We asked 21 researchers to individually annotate 10 papers. Each paper has been annotated by three annotators and was associated with $14 \pm 7.0$ topics using the majority voting strategy. The inter-annotator agreement was $0.45 \pm 0.18$ according to Fleiss’ Kappa, resulting in a moderate inter-rater agreement.

For the sake of space in Table 3 we report the values of precision, recall and f-measure of the top seven most performing approaches. The TF-IDF-M classifier ranks lists of words according to their TF-IDF score. These words are then mapped to CSO topics having Levenshtein similarity higher than 0.8. The LDA500-M classifier employs the Latent Dirichlet Allocation (LDA) [7] trained over the same set to find 500 topics. Similarly to TF-IDF-M, for a given paper, all the selected topics are mapped to the CSO topics. STM is the classifier originally adopted by Smart Topic Miner [19], the application used by Springer Nature for classifying proceedings within the Computer Science domain. It detects exact matches between the terms extracted from the text and the CSO topics. SYN represents the syntactic module of the CSO classifier, introduced in [24]. SEM consists of the semantic module of the CSO classifier. INT represents a hybrid version that returns the intersection of the topics produced by the SYN and SEM modules. Finally, CSO-C is the default implementation of the CSO Classifier which produces the union of the topics returned by the two modules. The overall values of precision and recall for a given classifier are computed as the average of the values of precision and recall obtained over the papers.

The LDA500-M and TF-IDF-M approaches performed poorly with an f-measure of 30.1%. STM and SYN yielded a very good precision of, respectively, 80.8% and 78.3%. These methods are able to find topics explicitly mentioned in
Table 3. Values of precision, recall, and f-measure. In bold the best results.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Description</th>
<th>Prec.</th>
<th>Rec.</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF-IDF-M</td>
<td>TF-IDF mapped to CSO concepts.</td>
<td>40.4%</td>
<td>24.1%</td>
<td>30.1%</td>
</tr>
<tr>
<td>LDA500-M</td>
<td>LDA with 500 topics mapped to CSO.</td>
<td>9.6%</td>
<td>21.2%</td>
<td>13.2%</td>
</tr>
<tr>
<td>STM</td>
<td>Classifier used by STM, introduced in [19].</td>
<td><strong>80.8%</strong></td>
<td>58.2%</td>
<td>67.6%</td>
</tr>
<tr>
<td>SYN</td>
<td>Syntactic module [24].</td>
<td>78.3%</td>
<td>63.8%</td>
<td>70.3%</td>
</tr>
<tr>
<td>SEM</td>
<td>Semantic module.</td>
<td>70.8%</td>
<td>72.2%</td>
<td>71.5%</td>
</tr>
<tr>
<td>INT</td>
<td>Intersection of SYN and SEM.</td>
<td>79.3%</td>
<td>59.1%</td>
<td>67.7%</td>
</tr>
<tr>
<td>CSO-C</td>
<td>The CSO Classifier [22].</td>
<td>73.0%</td>
<td><strong>75.3%</strong></td>
<td><strong>74.1%</strong></td>
</tr>
</tbody>
</table>

the text, which tend to be very relevant. However, they suffer from a low recall of 58.2% and 63.8% as they fail to identify more subtle topics. SEM has lower precision than SYN but higher recall and f-measure suggesting that it is able to identify further topics that do not directly appear in the paper. INT generated a higher precision (79.3%) compared to SYN and SEM (78.3% and 70.8%), but it did not yield a good recall dropping to 59.1%. Finally, CSO-C outperformed all the other methods in terms of both recall (75.3%) and f-measure (74.1%).

4.2 Academia/Industry Classification

In order to evaluate the quality of the academia/industry classification in AIDA we selected 100 papers: (i) 33 academic papers meaning that all the authors of each paper are reported with academic affiliations only; (ii) 33 industry papers whose related authors are reported with affiliation in industry only; (iii) 34 collaborative papers, meaning that each paper in this set includes authors with affiliations from academia and authors with affiliations from industry.

We then asked three independent researchers to manually annotate each paper as ‘academic’, ‘industry’ or ‘collaborative’ according to the classification above. The averaged agreement score of the three experts was 92.6%. We generated a gold standard by using a majority voting strategy. That is, if a paper was considered a academic paper by at least two researchers, it was labeled as such. There were not cases where a paper was annotated with three different classes by the researchers. We carried out a precision-recall analysis of the forecasted labels and obtained 100% as precision and recall.

4.3 Industrial Sector Classification

In order to evaluate the accuracy of our approach for identifying the industrial sectors of a document, we selected 100 organizations equally divided (20 per each industrial sector) among: telecommunication companies, healthcare companies, automotive companies, computing and information technology companies, and electronic companies. We then asked three independent experts to annotate each organization among the five classes above (or the other category if none of the previous categories was appropriate). The averaged agreement score of the experts was 84%.

We then created a gold standard using a majority voting strategy. For instance, if a company was classified as healthcare by at least two experts, then
its label was healthcare. To note that for each company at least two experts always gave the same label. We then performed a precision-recall analysis of the categories forecasted by our approach and, for each category, we obtained the accuracies shown in Table 4.

Table 4. Accuracy of the industry classification task

<table>
<thead>
<tr>
<th>Type</th>
<th>Automotive</th>
<th>Healthcare</th>
<th>Computing and IT</th>
<th>Electronic</th>
<th>Telecommunication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.947</td>
<td>0.947</td>
<td>0.864</td>
<td>0.756</td>
<td>0.918</td>
</tr>
</tbody>
</table>

5 Conclusions and Future Work

In this paper we have introduced AIDA, the Academic/Industry DynAmics Knowledge Graph. This resource characterizes 14M publications and 8M patents according to the research topics drawn from the Computer Science Ontology. 4M publications and 5M patents are also classified according to the type of the author’s affiliations and industrial sectors. In order to characterize documents according to their industrial sectors, we also designed the Industrial Sectors Ontology (INDUSO), which describes 66 sectors in a two level taxonomy.

AIDA was generated using an automatic pipeline that merges and integrates information from Microsoft Academic Graph, DBpedia, the Computer Science Ontology, and the Global Research Identifier Database. We evaluated different parts of the pipeline using a manually created gold standard obtained very competitive results.

AIDA allows researchers to analyse the evolution of research topics across academia and industry as well as to understand their dynamics within several industrial sectors. It can be used to identify the research trends of different industries and how and when academia and/or industry tackle these in particularly significant ways, thus facilitating a granular analysis of the interaction between these two worlds. Moreover, AIDA can also be employed to investigate authors, citations, countries and other entities already present in Microsoft Academic Graph.

The resource presented in this paper opens up several interesting directions of work. First, we will produce a comprehensive analysis of AIDA and the most significant research trends in academia and industry. We also intend to use AIDA to support systems for predicting the impact of specific areas of research on industry. Finally, we plan to explore the application of our pipeline to other fields, such as Biology and Engineering.

6 Acknowledgements

We gratefully acknowledge Springer Nature for funding this research. We thanks Dimensions for sharing their large dataset of patents.
References


23. Salatino, A.A., Thanapalasingam, T., Mannocci, A., Birukou, A., Osborne, F., Motta, E.: The computer science ontology: A comprehensive automatically-generated taxonomy of research areas. Data Intelligence 0(0), 1–38 (0). https://doi.org/10.1162/dint_a_00055


