UNICA

UNIVERSITÀ
DEGLI STUDI
DI CAGLIARI

Ph.D. DEGREE IN
Mathematics and Computer Science

Cycle XXXVI

TITLE OF THE Ph.D. THESIS

Academia/Industry DynAmics (AIDA): A knowledge Graph within the scholarly domain and its applications

Scientific Disciplinary Sector(s)

INF/01

Ph.D. Student:	Simone Angioni
Supervisor	Prof. Diego Reforgiato Recupero
Co-Supervisor	Prof. Roberto Tonelli

Final exam. Academic Year 2022/2023
Thesis defence: February 2024 Session

Università degli Studi di Cagliari

Dipartimento di Matematica e Informatica
Dottorato di Ricerca in Matematica e Informatica
Ciclo XXXVI

Ph.D. Thesis

Academia/Industry DynAmics (AIDA): A knowledge Graph within the scholarly domain and its applications

S.S.D. INF/01

Candidate
Simone Angioni

Supervisor
Prof. Diego Reforgiato Recupero

PhD Coordinator
Prof. Antonio Iannizzoto

Final examination academic year 2022/2023
February, 2024

Abstract

Scholarly knowledge graphs are a form of knowledge representation that aims to capture and organize the information and knowledge contained in scholarly publications, such as research papers, books, patents, and datasets. Scholarly knowledge graphs can provide a comprehensive and structured view of the scholarly domain, covering various aspects such as authors, affiliations, research topics, methods, results, citations, and impact. Scholarly knowledge graphs can enable various applications and services that can facilitate and enhance scholarly communication, such as information retrieval, data analysis, recommendation systems, semantic search, and knowledge discovery.

However, constructing and maintaining scholarly knowledge graphs is a challenging task that requires dealing with large-scale, heterogeneous, and dynamic data sources. Moreover, extracting and integrating the relevant information and knowledge from unstructured or semi-structured text is not trivial, as it involves natural language processing, machine learning, ontology engineering, and semantic web technologies. Furthermore, ensuring the quality and validity of the scholarly knowledge graphs is essential for their usability and reliability.

Contents

1 Introduction 13

2 Academia/Industry DynAmics (AIDA) Knowledge Graph 17

2.1 Introduction . . . 18
2.2 Background . . . 21
2.2.1 Scholarly Knowledge Graphs . . . 21
2.2.2 Relationship between Academia and Industry . . . 23
2.3 Approach . . . 23
2.4 Generation . . . 29
2.4.1 Topic Detection . . . 29
2.4.2 Integration of Affiliation Types . . . 32
2.4.3 Industrial Sector Classification . . . 32
2.5 AIDA Overview . . . 33
2.6 Sustainability . . . 36
2.7 Evaluation . . . 36
2.7.1 Topic Classification . . . 37
2.7.2 Academia/Industry and Industrial Sector Classifications . . . 39
2.7.3 Impact Forecasting . . . 41

3 AIDA Dashboard 45

3.1 Introduction . . . 46
3.2 Background . . . 47
3.2.1 Tools for assessing venues . . . 47
3.2.2 Other Scientometric tools . . . 48
3.3 Open Issues . . . 49
3.4 Dashboard . . . 50
3.4.1 The Back-end: Generation of the AIDA Dashboard Dataset . 50
3.4.2 The Web Interface . . . 52
3.4.3 Advanced Search . . . 56
3.5 Evaluation . . . 58
3.5.1 User Study . . . 58

4 AIDA-Bot **65**
4.1 Introduction . . . 66
4.2 Background . . . 67
4.3 AIDA-Bot . . . 70
4.3.1 Question Understanding . . . 71
4.3.2 Response Generator . . . 73
4.3.3 Transformer Models . . . 74
4.4 Evaluation . . . 75

5 Link Prediction **79**
5.1 Introduction . . . 80
5.2 Background . . . 82
5.2.1 Knowledge Graph Embeddings . . . 82
5.2.2 Review of State-of-the-art KGEs . . . 83
5.3 Open Issues . . . 84
5.4 Trans4E . . . 85
5.4.1 The Trans4E Model . . . 85
5.4.2 Link Prediction on N to M Relations . . . 86
5.5 Evaluation . . . 87
5.5.1 Evaluation Datasets . . . 87
5.5.2 Evaluation Criteria . . . 88
5.5.3 Hyperparameter Setting . . . 91
5.5.4 Results and Discussions . . . 92

6 AIDA-KG Analysis **97**
6.1 Introduction . . . 98
6.2 Open Issues . . . 98
6.3 Background . . . 99
6.4 Approach . . . 99
6.4.1 Data Selection . . . 99
6.4.2 Assessing Author Expertize . . . 100
6.4.3 Assessing Expertise Diversity in a Team of Authors . . . 100
6.4.4 Investigating the relationship between diversity of expertise and citations . . . 101
6.5 Results . . . 102
6.5.1 Max cosine distance between authors . . . 102
6.5.2 Number of components in the author graph . . . 103

7 Conclusion **107**

List of Figures **111**

Bibliography **113**

Acknowledgements

This thesis ends a significant chapter in my life. This journey that began when I entered bachelor degree to study computer science and has continued through the years. I tried to understand all aspects of this field, but I continue to find new things that make me ever more interested to learn and explore. Along the way, I have had the privilege to meet a lot of people from whom i learnt a lot. Their dedication to their work has left an indelible mark on my journey, and I am forever grateful for their influence.

Today, I recognize that there is still much more to discover and learn, and I eagerly anticipate the next steps in my academic and personal growth.

In this dissertation, I want to express my gratitude for the support and help I received while working on it. I'd like to start by thanking my supervisor, Prof. Diego Reforgiato Recupero. I'm truly thankful for his consistent support throughout the whole journey. It means a lot to have a supervisor who genuinely cared about me and my work, always ready to help with my questions and problems. I have learned a great deal under his mentorship, and his dedication to my progress is deeply appreciated.

I also wish to express my sincere appreciation to the supervisors I had during my international experiences, in chronological order: Dr. Angelo Salatino, Dr. Francesco Osborne, and Prof. Enrico Motta. I would also like to thank all the researcher group that hosted me, and all the people I have met in this adventure. All of you have made my abroad journey unforgettable.

Outside from the work, I would like to express my sincere gratitude to my parents, Paola and Roberto, and my brother Andrea, who I would like to thank with love and gratitude for all the support they provided me during this experience.

Lastly, I want to express my gratitude to my friends who have stood by my side, providing moments of happiness and friendship.

Publications

The research reported in this thesis has contributed to the following publications:

- [135] Salatino, A., Angioni, S., Osborne, F., Recupero, D. R., & Motta, E. (2023). Diversity of Expertise is Key to Scientific Impact: a Large-Scale Analysis in the Field of Computer Science. arXiv preprint arXiv:2306.15344

- [95] Meloni, A., Angioni, S., Salatino, A., Osborne, F., Recupero, D. R., & Motta, E. (2023). Integrating Conversational Agents and Knowledge Graphs Within the Scholarly Domain. IEEE Access, 11, 22468-22489.

- [17] Angioni, S., Salatino, A., Osborne, F., Birukou, A., Recupero, D. R., & Motta, E. (2022, October). Leveraging Knowledge Graph Technologies to Assess Journals and Conferences at Springer Nature. In International Semantic Web Conference (pp. 735-752). Cham: Springer International Publishing.

- [19] Angioni, S., Salatino, A., Osborne, F., Recupero, D. R., & Motta, E. (2022). The AIDA dashboard: a web application for assessing and comparing scientific conferences. IEEE Access, 10, 39471-39486.

- [18] Angioni, S., Salatino, A., Osborne, F., Recupero, D. R., & Motta, E. (2021). AIDA: A knowledge graph about research dynamics in academia and industry. Quantitative Science Studies, 2(4), 1356-1398.

- [103] Nayyeri, M., Cil, G. M., Vahdati, S., Osborne, F., Rahman, M., Angioni, S., ... & Lehmann, J. (2021). Trans4E: Link prediction on scholarly knowledge graphs. Neurocomputing, 461, 530-542.

- [96] Meloni, A., Angioni, S., Salatino, A., Osborne, F., Reforgiato Recupero, D., & Motta, E. (2021). Aida-bot: A conversational agent to explore scholarly knowledge graphs.

- [16] Angioni, S., Salatino, A., Osborne, F., Birukou, A., Recupero, D. R., & Motta, E. (2021). Assessing Scientific Conferences through Knowledge Graphs. In International Semantic Web Conference (ISWC) 2021: Posters, Demos, and Industry Tracks (Vol. 2980).

- [21] Angioni, S., Salatino, A. A., Osborne, F., Recupero, D. R., & Motta, E. (2020). Integrating knowledge graphs for analysing academia and industry dynamics. In ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium: International Workshops: DOING, MADEISD, SKG, BBIGAP, SIMPDA, AIMinScience 2020 and Doctoral Consortium, Lyon, France, August 25–27, 2020, Proceedings 24 (pp. 219-225). Springer International Publishing.

- [20] Angioni, S., Salatino, A., Osborne, F., Reforgiato Recupero, D., & Motta, E. (2020). The aida dashboard: Analysing conferences with semantic technologies.

- [15] Angioni, S., Osborne, F., Salatino, A., Reforgiato Recupero, D., & Motta, E. (2019). Integrating knowledge graphs for comparing the scientific output of academia and industry.

In addition, in my Ph.D. course I have worked in other projects contributing to the following publications:

- [14] Angioni, S., Lincoln-DeCusatis, N., Ibba, A., & Recupero, D. R. (2023). A transformers-based approach for fine and coarse-grained classification and generation of MIDI songs and soundtracks. PeerJ Computer Science, 9, e1410.
- [13] Angioni, S., Carta, S., Consoli, S., Reforgiato Recupero, D., & Stanciu, M. M. (2021, December). A Big Data framework based on Apache Spark for Industry-specific Lexicon Generation for Stock Market Prediction. In The 5th International Conference on Future Networks & Distributed Systems (pp. 616-624).
- [104] Nayyeri, M., Cil, G. M., Vahdati, S., Osborne, F., Kravchenko, A., Angioni, S., ... & Lehmann, J. (2021). Link prediction of weighted triples for knowledge graph completion within the scholarly domain. Ieee Access, 9, 116002-116014.

Chapter 1

Introduction

Scholarly knowledge graphs are a form of knowledge representation that aims to capture and organize the information and knowledge contained in scholarly publications, such as research papers, books, patents, and datasets. Scholarly knowledge graphs can provide a comprehensive and structured view of the scholarly domain, covering various aspects such as authors, affiliations, research topics, methods, results, citations, and impact. Scholarly knowledge graphs can enable various applications and services that can facilitate and enhance scholarly communication, such as information retrieval, data analysis, recommendation systems, semantic search, and knowledge discovery.

However, constructing and maintaining scholarly knowledge graphs is a challenging task that requires dealing with large-scale, heterogeneous, and dynamic data sources. Moreover, extracting and integrating the relevant information and knowledge from unstructured or semi-structured text is not trivial, as it involves natural language processing, machine learning, ontology engineering, and semantic web technologies. Furthermore, ensuring the quality and validity of the scholarly knowledge graphs is essential for their usability and reliability.

Research articles and patents are an ideal medium to analyze the knowledge generated and developed by academia and industry. Today, we have several large-scale knowledge graphs which describe research papers and patents according to their titles, abstracts, authors, organizations, and other metadata. However, these datasets cannot be directly used to analyze the research dynamics of academia and industry since they lack a high-quality characterization of the relevant research topics and industrial sectors. Current solutions for categorizing documents face several challenges. Firstly, they struggle to easily determine if a document originates from academia or industry. Secondly, their representation of research topics is often rudimentary, typically presenting them as a mere list of terms. Such an approach fails to differentiate research topics from other generic keywords, handle situations where the same research area might have multiple labels, and recognize the semantic relationships between research areas. An illustrative example is that documents discussing "Neural Network" should inherently be associated with broader concepts like "Machine Learning" and "Artificial Intelligence". Lastly, a significant limitation is that the prevailing datasets do not categorize companies based on their sectors. This omission hinders the ability to gauge the influence of a research topic on distinct industries. These issues also make it harder for machine learning tools, like neural networks, to predict research impacts and upcoming patents. Taking these limitations in mind, in this thesis we will focus about Academia/Industry DynAmics Knowledge Graph (AIDA-KG). AIDA-KG is a novel scholarly knowledge graph that associates papers and patents according to different representation in order to produce a knowledge graph usable for assessing the relationship between academia and industry.

Specifically, in this thesis we will investigates how the AIDA-KG is built, what are the challenges in building such a knowledge graph, the limitations, and we will investigates the principles, functionalities, and potential impact.

The research program addressed during the Ph.D. course was dedicated to the development of AIDA-KG, and on the development, analysis of tools and algorithms using this resource.

The main research questions addressed in the target domains are:

Q1. How to use Semantic Web and Machine Learning technologies to encapsulate various scholarly entities and produce a comprehensive scholarly Knowledge Graph?

Q2. How to use Semantic Web Technologies to generate and produce analytics for academic venues? What are the most interesting functionalities to support researchers and editors in analysing venues?

Q3. How can Artificial Intelligence and Semantic Web technologies be used to provide verified answers for conversational agents?

Q4. How can the embedding models fix the incompleteness of Scholarly Knowledge Graphs?

By answering to thee research questions, the main contributions provided in this research work are:

- A new knowledge graph for studying the research dynamics of academia and industry;
- A new web application for analyzing and assessing scholarly venues
- A new architecture to enhance Conversational Agents with Knowledge Graphs
- A new embedding model designed to provide link prediction for large-scale Knowledge Graphs
- A scientometric analysis to assess whether a diverse pool of expertize within a research team can influence their scientific impact

Chapter 2

Academia/Industry DynAmics (AIDA) Knowledge Graph

2.1 Introduction

Academia and industry share a complex, multifaceted, and symbiotic relationship. Their collaboration and exchange of ideas, resources, and persons [11] are conducive to the production of new knowledge that will ultimately shape the society of the future. Analyzing the knowledge flow between academia and industry, understanding which directions have the biggest potential, and discovering the best strategies to harmonize their efforts is thus a critical task for several stakeholders [136]. Governments and funding agencies need to regularly assess the potential impact of research areas and technologies to inform funding decisions. Commercial organizations have to monitor research developments and adapt to technological advancements. 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 analyzed from several perspectives in the literature, focusing for instance on the characteristics of direct collaborations [23], the influence of industrial trends on curricula [171], and the quality of the knowledge transfer [24]. However, most of the quantitative studies on this relationship were limited to small-scale datasets or focused on very specific research questions [32, 11].

Research articles and patents are an ideal medium to analyze the knowledge generated and developed by academia and industry [23, 24]. Today, we have several large-scale knowledge graphs which describe research papers according to their titles, abstracts, authors, organizations, and other metadata. Examples include Microsoft Academic Graph[1] [167], Scopus[2], Semantic Scholar[3], Aminer [179], CORE [79], OpenCitations [123], and others. Other resources, such as Dimensions[4], the United States Patent and Trademark Office (USPTO)[5], the Espacenet dataset[6], and the PatentScope corpus[7], offer a similar description of patents. However, these datasets cannot be directly used to analyze the research dynamics of academia and industry since they lack a high-quality characterization of the relevant research topics and industrial sectors.

There are three primary limitations observed in the current systems. Initially, the existing methods struggle to distinguish whether a document, be it a research paper or a patent, originates from the academic world or the corporate sector. Secondly, these systems generally provide a broad overview of research subjects, often merely presenting them as a list of terms either selected by the authors or derived from the abstract. Such an approach is inadequate [116] because it: i) cannot distinguish

[1] Microsoft Academic Graph - `http://aka.ms/microsoft-academic`

[2] Scopus - `https://www.scopus.com/`

[3] Semantic Scholar - `https://www.semanticscholar.org/`

[4] Dimensions - `https://www.dimensions.ai/`

[5] USPTO - `https://www.uspto.gov/`

[6] Espacenet dataset - `https://worldwide.espacenet.com/`

[7] PatentScope - `https://patentscope.wipo.int/`

research topics from other generic keywords; ii) cannot handle cases where multiple tags refer to an identical research domain; and iii) does not leverage the inherent semantic connections between various research fields. For example, it should be intuitive to deduce that documents labeled under the topic of Neural Networks also pertain to Machine Learning and Artificial Intelligence. This enhanced representation would enable us to fetch all papers related to the topic Artificial Intelligence, even if the metadata lacks the exact phrase "artificial intelligence". The third limitation is that contemporary academic datasets fail to categorize companies based on their industry sectors. As a result, assessing the influence of a particular subject (like sentiment analysis, deep learning, or the semantic web) on varied industries (such as automotive, finance, or energy) becomes challenging.

These limitations affect also the performance of machine learning systems, typically based on neural networks, for predicting the impact of research trends and forecasting patents [176, 48, 94, 127]. These solutions typically work with limited features, such as the number of patents associated with a topic for each year, since current datasets do not integrate articles and patents, lack a granular representation of research topics, and cannot distinguish whether a document was produced by academia or industry. We hypothesize that considering a richer characterization of this space would ultimately yield better performance in comparison to state-of-the-art approaches.

In this chapter, we introduce the Academia/Industry DynAmics (AIDA) Knowledge Graph, which describes 21M publications and 8M patents in the field of *Computer Science*. Papers and patents are associated to the research topics in the Computer Science Ontology (CSO). In addition, 5.1M publications and 5.6M patents are also characterized according to the type of the author's affiliations (e.g., academia, industry) and 66 industrial sectors (e.g., automotive, financial, energy, electronics) from the Industrial Sectors Ontology (INDUSO). AIDA is also linked to several other knowledge bases, including MAKG, Dimensions, Google Patents, GRID, DBpedia, and Wikidata.

AIDA is available at `http://w3id.org/aida/`. It can be downloaded as a dump or queried via a Virtuoso triplestore at `http://w3id.org/aida/sparql/`. We plan to release a new version of AIDA every six months, to regularly update the publications, the topics, and the industrial sectors.

AIDA was generated using an automatic pipeline that integrates data from Microsoft Academic Graph (MAG)[8], Dimensions, English DBpedia, the Computer Science Ontology (CSO), and the Global Research Identifier Database (GRID), respectively containing information about 242M research papers, 38M patents, 4.58M entities, 14K research topics, and 97K organizations.

The resulting knowledge base enables analyzing the evolution of research topics across academia and industry and studying the characteristics of several industrial sectors. For instance, it enables detecting the research trends most interesting for

[8] We used the dump released in April 2020.

the automotive sector or which prevalent industrial topics were recently adopted by academia. It can thus be utilized by a variety of deep learning methods for predicting the impact of research trends on industry and academia [176, 127, 49]. It can also be used to characterize authors, citations, countries, and several other entities in MAG according to their topics and industrial sectors. This makes it possible to study further dynamics such as the migration of researchers and the citation flow between academia and the industry.

We evaluated the different parts of the pipeline for generating AIDA on manually crafted gold standards yielding competitive results. We also report an evaluation of the impact of AIDA on forecasting systems for predicting the impact of research topics on the industry. Specifically, we tested five classifiers on 17 combinations of features and found that the forecaster based on Long Short-Term Memory neural networks and exploiting the full set of features from AIDA obtain significantly better performance ($p<0.0001$) than alternative methods.

A preliminary version of AIDA which included a smaller data set and a limited number of semantic relations was previously discussed in a short workshop paper [21]. The current paper greatly expands on that work by presenting a novel and up-to-date version of AIDA (including about 5M additional articles), an improved version of the pipeline for generating AIDA, a more extensive ontological schema, and a comprehensive evaluation of AIDA.

In summary, our main contributions include:

- the first official release of AIDA, a knowledge graph for studying the research dynamics of academia and industry;
- a pipeline for automatically generating AIDA based on a robust semantic model and a state-of-the-art topic detection approach;
- a detailed discussion of AIDA schema, content, and links to other knowledge graphs;
- an evaluation of the AIDA pipeline and its ability to classify documents in terms of research topics and industrial sectors;
- an illustrative overview of the Computer Science domain according to the data in AIDA.
- a discussion of AIDA possible usage that summarizes some research efforts that adopted preliminary versions of AIDA;
- an analysis of the current limitations of the AIDA pipeline and a sustainability plan developed in collaboration with Springer Nature for replacing MAG with a combination of Dimensions and DBLP, after MAG will be decommissioned at the end of 2021;

- an appendix detailing several exemplary SPARQL queries in order to support the reuse of AIDA.

The rest of the chapter is organized as follows. In Section 2.2, we review the literature on methods and datasets for studying and quantifying the relationship between academia and industry. Section 2.3 describes the approach we used to generate AIDA. In Section 2.4, we describe the pipeline to generate AIDA, give an overview of the resulting knowledge graph, and discuss our strategy for releasing new versions. Section 2.7 presents the evaluation of the different parts of the AIDA pipeline and the experiments showing that AIDA can support effectively deep learning approaches for predicting the impact of research topics. In Section 2.5 we focus on the usage of AIDA and report three exemplary research efforts that adopted preliminary versions of AIDA: i) a bibliometric analysis of the research dynamics across academia and industry, ii) a study of the main research trends in two main venues of Human-Computer Interaction

2.2 Background

2.2.1 Scholarly Knowledge Graphs

In recent years, there has been a conspicuous emergence of multiple knowledge graphs that integrates research publications and their associated metadata. One such example was the Microsoft Academic Graph (MAG) [167], deprecated since 2022, which was a heterogeneous knowledge graph available in RDF format, including metadata about 242 million scientific publications, such as citations, authors, institutions, journals, conferences, and fields of study. OpenAlex[9] is another scholarly knowledge graph, active since 2022, that has taken over the MAG project by integrating the original dataset and taking on the responsibility of keeping it updated. OpenAlex, like MAG, integrates metadata related to authors, institutions, journals, and fields of study. Another significant resource in this domain is the Semantic Scholar Open Research Corpus (ORC) [9], a dataset including approximately 185 million publications released by Semantic Scholar, an academic search engine provided by the Allen Institute for Artificial Intelligence (AI2). OpenCitations, an independent infrastructure organization for open scholarship, proposes the OpenCitations Corpus, which currently contains 55 million publications and 655 million citations [123]. 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, a repository of over 200 million publications generated and utilized by the AMiner [179] system, is another noteworthy resource. AMiner is a free online academic search and mining system that extracts researchers' profiles from the web and integrates them into the metadata.

[9] `https://openalex.org`

CORE, a repository containing 24 million open access research outputs from repositories and journals worldwide, is a valuable resource in this context (CORE, 2011). The Dimensions corpus, produced by Digital Science, integrates and interlinks 109 million research publications, 5.3 million grants, and 40 million patents. Publications and citations from Dimensions are freely available for personal, non-commercial use.

Similarly, there are several resources focusing specifically on patents [146]. For instance, the European Patent Office (EPO) curates the Espacenet dataset, which currently encompasses about 110 million patents from around the world. Similarly, the United States Patent and Trademark Office produces a corpus that includes more than 14 million US patents. The World Intellectual Property Organization (WIPO) offers the PatentScope dataset, containing 84 million patent documents, including 4 million international patent applications.

Another category of scientific knowledge graphs includes those that incorporate a semantic representation of the content found within scientific articles. The Semantic Web community has been working for a while on this direction, promoting the Semantic Publishing paradigm [148], creating bibliographic repositories in the Linked Data Cloud [111], generating knowledge bases of biological data [29], formalising research workflows [172], extracting knowledge graphs from research papers [116, 41], implementing systems for managing nano-publications [63, 80] and micropublications [145], and developing a variety of ontologies to describe scholarly data, e.g., SWRC[10], BIBO[11], BiDO[12], FABIO[13], SPAR[14] [122], CSO[15] [142], and SKGO[16] [57]. Several other knowledge bases describe the research areas of scientific publications, such as the Medical Subject Heading (MeSH)[17] in Biology, Mathematics Subject Classification (MSC)[18] in Mathematics, Physics Subject Headings (PhySH)[19] in Physics, and many others.

In Computer Science, there are have several taxonomies of research areas. The best-known are the ACM Computing Classification System[20] and the Computer Science Ontology (CSO) [142]. 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.

[10] SWRC - http://ontoware.org/swrc

[11] BIBO - http://bibliontology.com

[12] BiDO - http://purl.org/spar/bido

[13] FABIO - http://purl.org/spar/fabio

[14] SPAR - http://www.sparontologies.net/

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

[16] SKGO - https://github.com/saidfathalla/Science-knowledge-graph-ontologies

[17] Medical Subject Heading - https://www.ncbi.nlm.nih.gov/mesh

[18] Mathematics Subject Classification - https://mathscinet.ams.org/msc

[19] Physics Subject Headings - https://physh.aps.org/

[20] ACM Classification System - https://www.acm.org/publications/class-2012

Currently, there are no datasets that enable the study of high-quality research topics across research papers and patents. For this reason, we decided to undertake this new endeavor and develop AIDA.

2.2.2 Relationship between Academia and Industry

Academia and industry typically tend to influence each other by exchanging ideas, resources, and researchers [124]. Analyzing their relationship, how they interact, and how research flows, allows us to understand their role within the whole knowledge economy [12]: from production, towards adoption, enrichment, and ultimately deployment as a new commercial product or service. In some cases, academia and industry engage in collaborations as an opportunity for a more productive division of tasks: academia focusing on scientific insights, and industry on commercialization [32]. Jack Stilgoe in his "*Who's driving innovation? New technologies and the collaborative state*" book [154] discusses the main drivers of scientific innovation and focuses on the central role of the industry sector in pushing innovation by constantly deploying new technologies. However, it can be argued that innovation advances also through a more complex route, which involves the birth of a new scientific area, the development of its theoretical framework, and the creation of innovative products that capitalize on the new knowledge [81].

In literature we can find some approaches aimed at studying the relationship between academia and industry, using both qualitative and quantitative approaches. One qualitative study is from [97] who share their personal experience on how the collaboration between industry and academia impacted their research program. Similarly, [62] performed a survey-based analysis to understand the innovation performance associated with collaborations between universities and German manufacturers. We can also find more quantitative approaches, such as [84], who employed both research papers and patents to understand the primary interests of both sides in this symbiosis. [69] analysed 20K research papers and 8K patents in the area of *fuel cells*, to assess the direct benefits for both academia and industry when they engage in a collaboration.

However, all of these approaches either focus on relatively narrow areas of science or are restricted to a limited number of research questions. By developing AIDA we are opening up new lines of inquiry for all practitioners that are interested in investigating the relationship between academia and industry and predicting their trends.

2.3 Approach

The Academia/Industry DynAmics (AIDA) Knowledge Graph includes about 1.3B triples that describe a large collection of publications and patents in *Computer Science* according to their research topics, industrial sectors, and author's affiliations (academia, industry, or collaborative). Specifically, 21M publications from MAG and 8M patents from Dimensions are classified 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[21].

The 5.1M publications and 5.6M patents that were associated with GRID IDs in the original data are also classified according to the type of the author's affiliations (e.g., academia, industry) and 66 industrial sectors (e.g., automotive, financial, energy, electronics) drawn from the Industrial Sectors ontology (INDUSO)[22], which was specifically designed to support AIDA.

Since these annotations require at least an affiliation of the authors of the document to be associated with a GRID ID (as detailed in Section 2.4), they are currently restricted only to the document linked to GRID by Microsoft Academics Graph and Dimensions.

About 4.5M articles and 4.9M patents were also typed with the three main categories of our schema: academia, industry, and collaboration (between academia and industry). We also included additional affiliation categories from GRID, such as "Government", "Facility", "Healthcare", and "Nonprofit".

AIDA was generated 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 2.1: AIDA - Affiliation Types.

	Publications	**Patents**
Academia	3,906,131	122,390
Industry	834,443	4,760,614
Collaborative	133,781	16,806
Additional categories in GRID	627,179	747,618
Documents with GRID ID	5,133,171	5,639,252
Total documents	20,850,710	7,940,034

Table 2.1 shows the number of publications and patents from academia, industry, and collaborative efforts. Please note that only the documents associated with a GRID ID (about 5.1M publications and 5.6M patents) can be classified as academia, industry, collaborative or any other additional category from GRID.

When considering the affiliation types, most publications (69.8%) are written by academic institutions, however, the industry contributes to a good number of them (15.3%). The situation is reversed when considering patents: 84% of them are from industry and only 2.3% from academia. Another interesting finding is that the

[21] With $x \pm y$ we refer to x being the average and y the standard deviation.

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

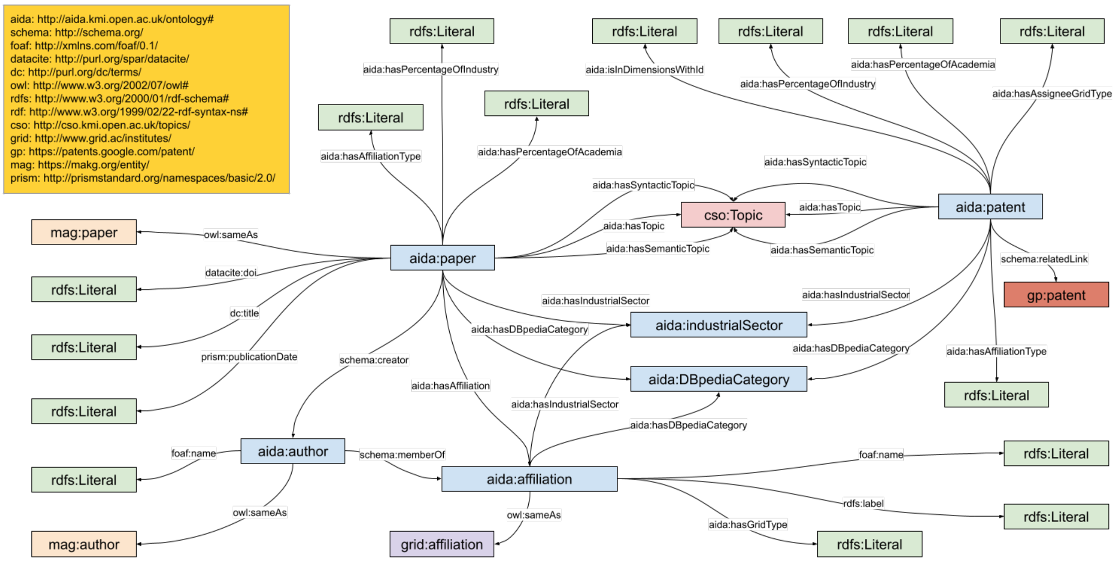


Figure 2.1: AIDA KG data model. For an enlarged version, please visit `http://w3id.org/aida#aidaschema`.

collaborative efforts are limited, involving only 2.6% of the publications and 0.2% 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 AIDA Schema, Schema.org, FOAF, OWL, CSO and others. We created AIDA Schema to define all the specific relations that could not be reused from state-of-the-art ontologies. It is available at `http://w3id.org/aida/ontology`.

Figure 2.1 depicts the full data model of AIDA KG, including both relations that we defined within AIDA schema and the ones we imported from external schemas. It focuses on six types of entities (light-blue boxes in Figure 2.1): papers, patents, authors, affiliations, industrial sectors, and DBpedia categories. To be compatible with other knowledge graphs in this space (e.g., MAG, Scopus, DBLP, Semantic Scholar), papers are identified according to their Digital Object Identifier (DOI) and patents according to their World Intellectual Property Organization (WIPO) ID. We also retain the original MAG IDs for papers and authors as additional identifiers. These are used to link AIDA to MAKG and to identify articles that lack a DOI. In addition, affiliations are identified with GRID IDs. Industrial sectors and DBpedia categories are identified according to the instances available within INDUSO.

The main information about papers and patents are given by means of the following semantic relations:

- *hasTopic*, which associates with the documents all their relevant topics drawn from CSO.

- *hasIndustrialSector*, which associates with documents and affiliations the relevant industrial sectors drawn from INDUSO.

- *hasAffiliationType*, which associates with the documents the three categories (academia, industry, or collaborative) describing the affiliations of their authors.

AIDA schema includes also some additional relationships which support more complex queries:

- *hasSyntacticTopic* and *hasSemanticTopic*, which indicate, respectively, all the topics extracted using the syntactic module and the semantic module of the CSO Classifier [138]. The first set is composed by topics that are explicitly mentioned in the documents. It has high precision but low recall and may be used by applications for which precision is paramount. The second one consists of topics that do not directly appear in the text but were inferred using word embeddings.

- *hasAffiliation*, which identifies the affiliations of a paper.

- *hasPercentageOfAcademia* and *hasPercentageOfIndustry*, which link to articles and patents the percentage of authors from academia and industry. It may be used to generate analytics that need to further segment the *collaborative* category.

- *hasGridType*, *hasAssigneeGridType*, which associate the eight categories of organizations described in GRID (Education, Healthcare, Company, Archive, Nonprofit, Government, Facility, and Other) with affiliations and patents.

- *hasDBpediaCategory*, which associates with papers the industrial categories found in DBpedia (through the *About:Property* and *About:Industry*).

- *isInDimensionsWithId*, which identifies the patent id used within the Dimensions database.

As already mentioned, the AIDA knowledge graph also adopts several relations from external sources. These are:

- *https://schema.org/creator*, which links documents to authors and authors to affiliations.

- *https://schema.org/memberOf*, which links authors to affiliations.

- *http://www.w3.org/1999/02/22-rdf-syntax-ns#type*, which defines the type of the entity.

- *http://www.w3.org/2000/01/rdf-schema#label*, which indicates the label of an affiliation.

Table 2.2: Number of triples for each relation in AIDA

Provenance	Relation	N. Triples
AIDA	http://aida.kmi.open.ac.uk/ontology#hasTopic	847,931,791
AIDA	http://aida.kmi.open.ac.uk/ontology#hasSemanticTopic	159,711,581
AIDA	http://aida.kmi.open.ac.uk/ontology#hasSyntacticTopic	70,349,962
AIDA	http://www.w3.org/2000/01/rdf-schema#type	54,839,960
AIDA	http://www.w3.org/2002/07/owl#sameAs	46,950,925
AIDA	http://aida.kmi.open.ac.uk/ontology#hasIndustrialSector	12,006,596
AIDA	http://aida.kmi.open.ac.uk/ontology#hasAffiliationType	9,774,165
AIDA	http://aida.kmi.open.ac.uk/ontology#hasDBpediaCategory	9,691,511
AIDA	http://aida.kmi.open.ac.uk/ontology#isInDimensionWithId	7,940,034
AIDA	http://schema.org/relatedLink	7,940,034
AIDA	http://aida.kmi.open.ac.uk/ontology#hasPercentageOfAcademia	4,179,108
AIDA	http://aida.kmi.open.ac.uk/ontology#hasPercentageOfIndustry	5,745,644
MAG	http://schema.org/creator	53,647,155
MAG	http://xmlns.com/foaf/0.1/name	26,048,450
MAG	http://purl.org/dc/terms/title	20,850,710
MAG	http://prismstandard.org/namespaces/basic/2.0/publicationDate	20,850,710
MAG	http://purl.org/spar/datacite/doi	5,636,401
MAG	http://schema.org/memberOf	4,828,260
MAG	http://aida.kmi.open.ac.uk/ontology#hasAffiliation	6,613,216
GRID	http://aida.kmi.open.ac.uk/ontology#hasAssigneeGridType	5,056,426
GRID	http://aida.kmi.open.ac.uk/ontology#hasGridType	13,171
GRID	http://www.w3.org/2000/01/rdf-schema#label	13,171

- *http://purl.org/dc/terms/title*, which indicates the title of a paper.
- *http://purl.org/spar/datacite/doi*, which indicates the DOI of a paper.
- *http://xmlns.com/foaf/0.1/name*, which indicates the name of an author or an affiliation.
- *http://schema.org/relatedLink*, which states the related link of a patent (typically a Google Patent URL).
- *http://prismstandard.org/namespaces/basic/2.0/publicationDate*, which indicates the year of publication of a paper.
- *http://www.w3.org/2002/07/owl/sameAs*, which links papers, authors, or affiliations to their representations on external knowledge bases.

Table 2.2 reports the number of triples available in the current version of AIDA for each relation. AIDA includes a total of about 1.3B triples: 1.2B with object properties and 98M with datatype properties. Here, we distinguish the provenance of the triples to highlight which ones are directly generated by the AIDA pipeline (described in Section 2.4) and which ones are reused from other knowledge graphs. Overall, 1.18B triples (89,1 % of the total) were generated by our pipeline, while 185M were derived from MAG and 7M from GRID. We reused some relations from MAG, because they enable several kinds of useful queries involving, for instance, the years of publication of the articles and the names of the authors. In the set of triples generated by the AIDA pipeline, 1.08B (82,6%) regard the three main

Table 2.3: Links of AIDA with external Knowledge Bases.

Knowledge Base	Type	Distinct Entities	Total triples
CSO	Topic	11,091	1,077,993,334
MAKG	Author	26,035,279	26,035,279
MAKG	Paper	20,850,710	20,850,710
INDUSO	Industrial Sector	66	12,007,438
Dimensions	Patent	7,940,034	7,940,034
Google Patents	Patent	7,940,034	7,940,034
GRID	Affiliation	13,171	13,171
DBpedia	Organization	13,171	13,171
DBpedia	Concept	3,864	3,864
Wikidata	Concept	3,842	3,842

contributions of AIDA. Specifically, 1.07B triples regard the topics (*hasSyntacticTopic*, *hasSemanticTopic*, *hasTopic*), 19,6M the affiliation types (*hasAffiliationType*, *hasPercentageOfAcademia*, *hasPercentageOfIndustry*), and 12.0M the industrial sectors (*hasIndustrialSector*).

Table 2.3 reports the number of triples linking AIDA to external knowledge bases and the number of relevant distinct entities. For instance, AIDA includes more then 1B triples having as object a topic in CSO and overall links to 11K unique topics. AIDA is mostly linked to MAKG (the RDF version of MAG), including *own:sameAs* relationships for 21M papers and 25M authors. It also links to Dimensions (8M patents), Google Patents (8M patents), GRID (13K affiliations), and DBpedia (3,864 concepts and 13K affiliations), and Wikidata (3,842 concepts). It should be noted that we cannot link directly to MAG, since it is not available online. However, since we use MAG IDs for papers and authors, mapping MAG and AIDA is trivial.

AIDA includes also the most recent mappings between CSO and DBpedia and between CSO and Wikidata, which implicitly links the documents in AIDA to 3,864 DBpedia entities and 3,842 Wikidata entities. Currently, those statements are not materialized for reason of space. However, materializing these links would yield additional 460M triples linking papers and patents to DBpedia entities (e.g., `http://dbpedia.org/resource/Machine_learning`) and 450M triples linking them to Wikidata entities (e.g., `http://www.wikidata.org/entity/Q2539`). Alternatively, the user can explore these links by formulating SPARQL queries that take advantage of the *owl:sameAs* relationship between CSO, DBpedia, and Wikidata (see example in the Appendix).

The online documentation of AIDA schema is available at `https://w3id.org/aida#aidaschema`.

AIDA is accessible via a Virtuoso triplestore at `http://w3id.org/aida/sparql`. The user can click the "help" button in the upper right of the web page for instructions on how to use the endpoint and some exemplary queries. The full dump of the last versions of AIDA is available at `http://w3id.org/aida/`. The dumps of the previous versions are available at `http://w3id.org/aida/downloads.php#datasets`.

AIDA is licensed under a Creative Commons Attribution 4.0 International Li-

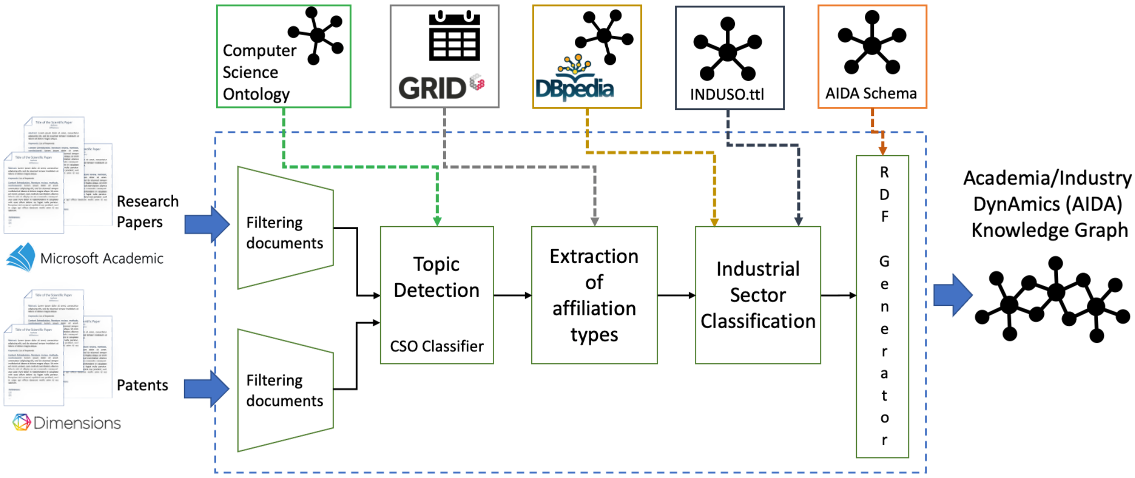


Figure 2.2: Workflow for the generation of AIDA.

cense (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 chapters, we will describe the pipeline for the automatic generation of AIDA (Section 2.4) and present an overview of the data (Section 2.5).

2.4 Generation

The automatic pipeline for generating AIDA works in three steps: topics detection, integration of affiliation types, and industrial sector classification, as shown in Figure 2.2.

In the following, we will describe each phase of the process (Paragraph 2.4.1 - 2.4.3).

2.4.1 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) [150], 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).

In the current version, the resulting dataset includes 21M publications and 8M patents. The publications (21M) and authors (25M) extracted from MAG are also linked (*owl:sameAs*) to the relevant entities in MAKG. The patents obtained from Dimensions (8M) are linked (*schema:relatedLink*) to the relevant patents in Google 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 each document with the research topics from the Computer Science Ontology (CSO) [142]. CSO is an automatically generated ontology of research topics in the field of *Computer Science*. We used the current version (3.2), which includes 14K research topics and 159K semantic relationships. The CSO data model[23] is an extension of SKOS[24] and the main semantic relationships are *superTopicOf*, which is used to define the hierarchical relations within the field of Computer Science (e.g., <*artificial intelligence, superTopicOf, machine learning*>) and *relatedEquivalent*, which is used to define alternative labels for the same topic (e.g., <*ontology matching, relatedEquivalent, ontology alignment*>).

We adopted CSO since it offers a much more granular characterization of research topics than standard classification schemas (e.g., the ACM Classification) and generic knowledge graphs (e.g., DBpedia, Wikidata). For instance, a recent analysis [140] reported that less than 37% of the topics in CSO are covered by DBpedia.

CSO was officially released in 2019 and has been already adopted by several major organizations, including Springer Nature. In the last two year, CSO supported the creation of many innovative applications and technologies, including ontology-driven topic models (e.g., CoCoNoW [28]), recommender systems for articles (e.g., SBR [156]) and video lessons [37], visualisation frameworks (e.g., ScholarLensViz [89], ConceptScope [178]), temporal knowledge graphs (e.g., TGK [133]), NLP frameworks for entity extraction [53], tools for identifying domain experts (e.g., VeTo [165]), and systems for predicting academic impact (e.g., ArtSim [44]). It was also used for several large-scale analyses of the literature (e.g., Cloud Computing [91], Software Engineering [47], Ecuadorian publications [47]).

We annotated publications and patents using the CSO Classifier [138], an open-source Python tool[25] that we developed for annotating documents with research topics from CSO [139].

The CSO Classifier was initially developed in the context of a collaboration with Springer Nature, with the aim of automatically classifying scientific volumes

[23] CSO Schema - `https://cso.kmi.open.ac.uk/schema/cso`

[24] Simple Knowledge Organization System - `https://www.w3.org/2004/02/skos/`

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

according to a granular set of research areas. In this context, it supported Smart Topic Miner [137], a web application for assisting the Springer Nature editorial team in annotating conference proceedings in Computer Science, such as LNCS, LNBIP, CCIS, IFIP-AICT and LNICST. This solution brought a 75% cost reduction and dramatically improved the quality of the annotations, resulting in 12M additional downloads over 3 years from the SpringerLink portal[26].

The CSO Classifier is an unsupervised method that operates in three phases. First the syntactic module finds all topics in the ontology that are explicitly mentioned in the paper. Secondly, a semantic module identifies further semantically related topics using part-of-speech tagging and similarity over word embeddings. Finally, the CSO Classifier enriches the resulting set by including the super-areas of these topics according to CSO.

Specifically, in the *syntactic* module, the text is split into unigrams, bigrams, and trigrams. Each n-gram is then compared with concepts labels in CSO using the Levenshtein similarity. As result, it returns all matched topics having similarities greater than or equal to the pre-defined threshold.

The *semantic* module takes advantage of a pre-trained Word2Vec word embedding model which captures semantic properties of words [98]. We trained this model using titles and abstracts of over 4.6M English publications in the field of Computer Science from MAG. We pre-processed this data by replacing spaces with underscores in all n-grams matching the CSO topic labels (e.g., “semantic web” became “semantic_web”). We performed also a collocation analysis to identify frequent bigrams and trigrams (e.g., “highest_accuracies”, “highly_cited_journals”). This solution allows the CSO Classifier to better disambiguate concepts and treat terms such as “deep_learning” and “e-learning” as completely different words. The model parameters are: *method* = skipgram, *embedding-size* = 128, *window-size* = 10, *min-count-cutoff* = 10, *max-iterations* = 5. The semantic module based on these embeddings identifies candidate terms composed of a combination of nouns and adjectives using a part-of-speech tagger. Then, it splits these candidate terms into unigrams, bigrams, and trigrams. For each n-gram we retrieve its most similar word from the Word2Vec model and we compute their cosine similarity with the topic labels in CSO. For bigrams and trigrams, we firstly check in the model their glued version, creating one single word, e.g., “semantic_web”. If this word is not available within the model vocabulary, the classifier uses the average of the embedding vectors of all its tokens. Then, for each identified topic, the CSO Classifier computes the relevance score as the product between the number of times it was identified (frequency) and the number of unique n-grams that helped it to be inferred (diversity). Finally, it uses the elbow method [144] for selecting the set of most relevant topics.

Finally, the resulting set of topics is enriched by including all their super-topics in CSO up to the root: *Computer Science*. For instance, a paper tagged as *neural network* is also tagged with *machine learning*, *artificial intelligence* and *computer*

[26] SpringerLink - `https://link.springer.com/`

science. This solution yields an improved characterization of high-level topics that are not directly referred to in the documents.

The reader notices that the CSO ontology contains nine levels of topics. When we detect a specific topic (e.g., Neural Networks) we also infer all the super topics in the CSO taxonomy (Machine Learning, Artificial Intelligence, Computer Science). The user can choose to just use the topics directly mentioned in the paper (*hasSyntacticTopic*), the ones inferred by using word embeddings (*hasSemanticTopic*), or the full set of topics that also includes the super-topics (*hasTopic*). More details about the CSO Classifier are available in [138].

We also import in AIDA the mapping between CSO and DBpedia, which is a set of 3,864 *owl:sameAs* relationships aligning the two knowledge bases and the mapping between CSO and Wikidata, which includes 3,842 *owl:sameAs* relationships. This allows us to establish several implicit links between documents in AIDA and concepts in DBpedia and Wikidata, which can be materialized with a reasoner or queried using SPARQL (see example in the Appendix).

2.4.2 Integration 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 IDs. 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). In total 5.1M articles and 5.6M patents were associated with GRID IDs. We leverage this last field to tag 4.5M articles and 4.9M patents as ‘academia’, ‘industry’, or ‘collaborative’. A document is assigned an ‘academia’ type if all the authors or original assignees have an academic affiliation (’Education’ in GRID), an ‘industry’ type if they have an industrial affiliation (’Company’ in GRID), 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*.

2.4.3 Industrial Sector Classification

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 into 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 set of 699 sectors. We then applied a bottom-up hierarchical clustering approach for merging similar sectors. For instance, the industrial sector “Computing and IT" was derived from categories such as “Networking hardware", “Cloud Computing", and “IT service management".

This structure was used as a starting point by a team of ontology engineers from the Open University and the University of Cagliari and domain experts from Springer Nature, who manually revised these categories and arranged the resulting sectors in a two-level taxonomy.

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, which links the first level sectors to the second level sectors.
- the *prov:wasDerivedFrom* property, which associates each of the 66 industrial sectors to the original 699 sectors that were derived from DBpedia.
- the *rdf:type* property, which is used to define the 66 sectors as *:industrialSector* and the original 699 sectors as *:DBpediaCategory*

To tag a document with INDUSO, we identify its affiliations on DBpedia using the link between GRID and DBpedia and then retrieve the objects of the properties *About:Purpose* and *About:Industry.* We then use the previously defined mapping between DBpedia and INDUSO to obtain the industrial sectors.

For instance, a document with an author affiliation described in DBpedia as ‘natural gas utility’ is tagged with the second level sector ‘Oil and Gas Industry’ and the first level sector ‘Energy’.

2.5 AIDA Overview

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

Figure 2.3 shows the 16 high-level topics (direct sub-topics of Computer Science in CSO) associated with most research articles in AIDA and reports the relevant percentage of academic publications, industrial publications, academic patents, and industrial patents.

These figures were computed by normalizing the number of documents associated with a topic in a category (e.g., academic publications) with the total number of documents in the same category. It should be noted that the percentages do not add to 100% since documents can be associated with multiples topics.

Some topics, such as Artificial Intelligence and Theoretical Computer Science, are mostly addressed by academic publications. Other ones, e.g., Computer Security, Computer Hardware, and Information Retrieval attract a stronger interest from the industry. The topics which are mostly associated with patents are Computer Networks, Internet, and Computer Hardware.

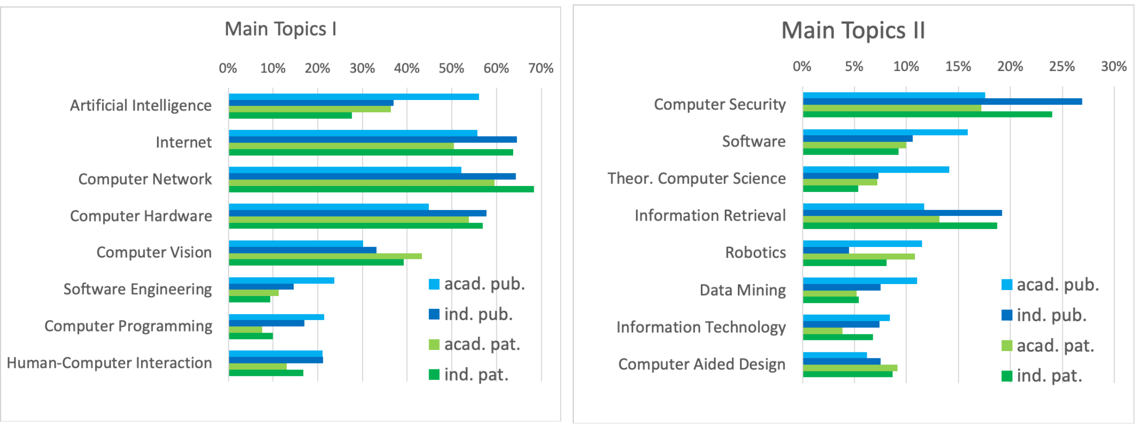


Figure 2.3: Distribution of the main topics.

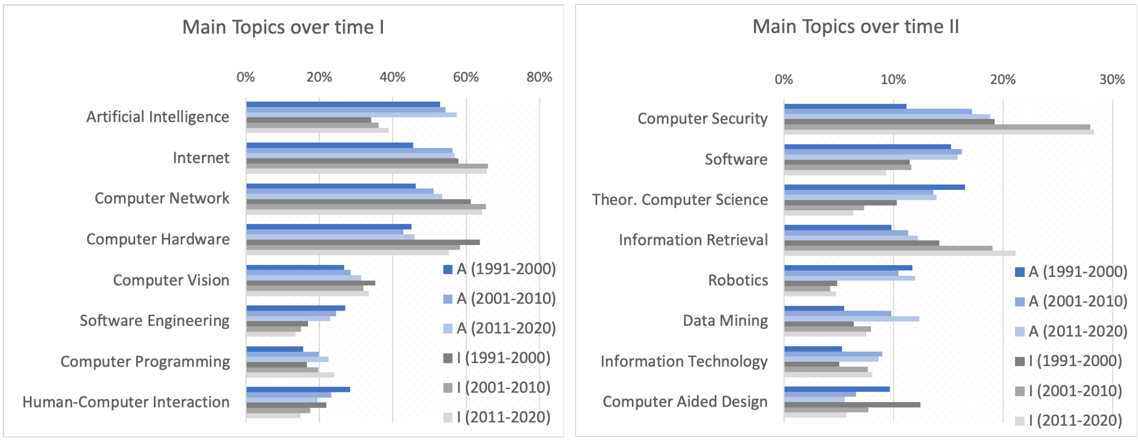


Figure 2.4: Distribution of the topics in publications across time.

Figure 2.4 shows the percentage of publications from academia (A) and industry (I) for the same 16 topics across three windows of time (1991-2000, 2001-2010, and 2011-2020). The split in three intervals of ten years is useful to highlight the trend of each topic across the years.

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.

Figure 2.5 (Main Industrial Sectors I and Main Industrial Sectors II) shows the 16 industrial sectors associated with most research articles and reports their percentage of publications and patents in AIDA.

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,

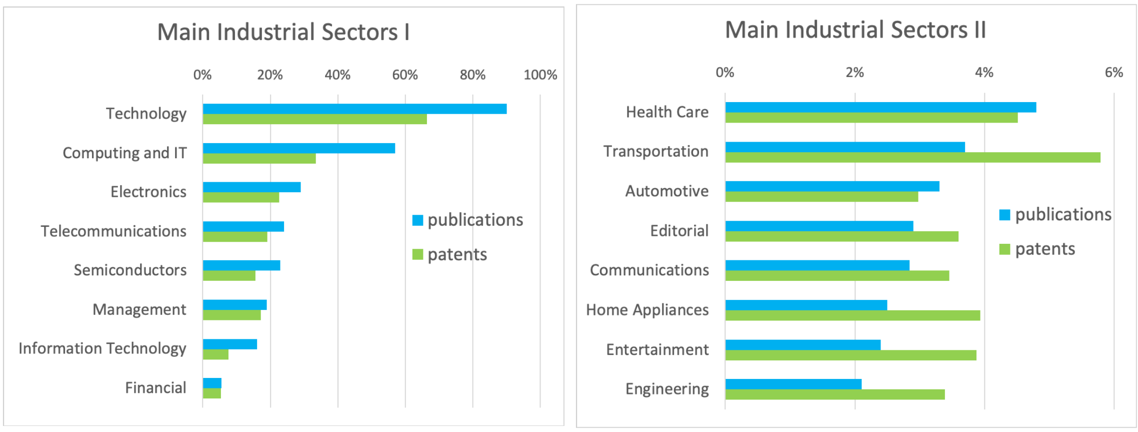


Figure 2.5: Distribution of the main industrial sectors.

and Editorial.

AIDA also enables to analyze how these sectors have a different composition in regards to research topics. Table 2.4 highlights the key topics of a set of exemplary sectors by reporting the difference between the normalized 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.

Table 2.4: Topic composition of some prominent industrial sectors. In bold the highest value for each row.

	Computing and IT	Telecommunications	Electronics	Semiconductor	Inf. Technoloogy	Photograpy	Automotive	Financial
Artificial Intelligence	9%	5%	9%	-17%	0%	22%	8%	-6%
Computer Aided Design	-21%	-27%	-2%	**90%**	1%	-5%	2%	-36%
Computer Hardware	-7%	7%	-7%	31%	-5%	-12%	-9%	-17%
Computer Network	-3%	**17%**	-9%	11%	-9%	-18%	-15%	-8%
Computer Programming	18%	-19%	-1%	12%	52%	-31%	-16%	-32%
Computer Security	6%	-1%	-2%	-27%	-1%	9%	-35%	21%
Computer Systems	1%	1%	-3%	1%	4%	-2%	-12%	-10%
Computer Vision	-7%	-1%	**21%**	-16%	-29%	44%	-7%	**52%**
Data Mining	**28%**	-25%	12%	-35%	49%	-18%	-34%	-17%
Human-computer Inter.	14%	-9%	8%	-41%	9%	-21%	-6%	32%
Information Retrieval	6%	-16%	14%	-55%	-6%	**71%**	-37%	29%
Information Technology	20%	-15%	-5%	-41%	55%	13%	-41%	-20%
Internet	4%	13%	-6%	-1%	1%	-19%	-24%	-4%
Operating Systems	14%	-40%	-8%	1%	**61%**	-24%	-55%	-30%
Robotics	3%	-1%	16%	-14%	-9%	-18%	**322%**	15%
Software Engineering	22%	16%	6%	2%	55%	-24%	20%	-31%

The industrial sectors have a very distinct composition, even when considering just the high-level topics in the table. For instance, the Automotive sector focuses 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.

AIDA can also be queried via triplestore using SPARQL[27]. The ontological schema of AIDA allows users to formulate queries about topics, industrial sectors, and affiliation types associated with articles and patents. In the Appendix of this manuscript we report a selection of sample queries that can be run on our SPARQL endpoint.

2.6 Sustainability

We plan to keep maintaining and updating the resource in the following years. For this reason, we set up an automatic pipeline that will update the data every 6 months.

At the end of 2021 Microsoft decommissioned the MAG project[28]. We thus decided to introduce two additional datasets within our integration pipeline: OpenAlex[29] and DBLP[30]. We included OpenAlex because it shares the same schema with MAG and it has a low cost of integration. However, since OpenAlex does not disambiguate conferences yet, we leveraged the conference representation of DBLP, by mapping papers across the two datasets. To achieve this, we designed a two-stage pipeline. We firstly mapped papers with the same DOI. Then, for the conferences that do not assign DOIs to articles (e.g., AAAI, NeurIPS), we mapped the papers across the two datasets by computing the string similarity of their titles. Future versions of AIDA KG and the generated analytics will be based on these newly integrated datasets.

2.7 Evaluation

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

[27] AIDA triplestore - `http://w3id.org/aida/sparql`

[28] Next Steps for Microsoft Academic – Expanding into New Horizons - `https://www.microsoft.com/en-us/research/project/academic/articles/microsoft-academic-to-expand-horizons-with-community-driven-approach/`

[29] OpenAlex - `https://openalex.org`

[30] DBLP - `https://dblp.org`

2.7.1 Topic Classification

We compared the CSO Classifier, which we use to annotate documents according to their topics, against thirteen unsupervised approaches using a gold standard made of 70 most cited papers [138] within the fields of Natural Language Processing (23 papers), Semantic Web (23), and Data Mining (24). We chose the most cited papers since this solution offers a simple, deterministic, and not arbitrary selection criteria. The 70 papers were annotated by 21 human experts. Each human expert annotated 10 papers; each paper was annotated by 3 human experts resulting in 210 annotations overall. The 21 experts were researchers working in different areas of Computer Science with over 5 years of experience. They were asked to read title, abstract and keywords and assign all the relevant topics from the CSO ontology so as to emulate the classifier's task. Each paper was associated with 14 ± 7.0 topics using the majority voting strategy.

The inter-annotator agreement was 0.45 ± 0.18 according to Fleiss' Kappa, resulting in a moderate inter-rater agreement.

It should be noted that this range of agreement is normal when using a large number of granular categories, such as the 14K topics in CSO.

In Table 2.5 we report the values of precision, recall, and F1 of all tested classifiers.

The first eight classifiers are based on TF-IDF and Latent Dirichlet Allocation (LDA) [33], and their performance did not exceed a F1 of 30.1%. For each paper, TF-IDF returns a ranked list of words according to their TF-IDF score. The TF-IDF-M classifier, instead, returns the set of CSO topics having Levenshtein similarity higher than 0.8 with the words with the best TF-IDF score. This threshold was set empirically, because it yielded the best performance for the baselines.

LDA100, LDA500, LDA1000 are three LDA classifiers, respectively trained on 100, 500 and 1000 topics. These three classifiers select all LDA topics with a probability of at least j and return all their words with a probability of at least k. The best values of j and k were found performing a grid search. In a similar way, we trained LDA100-M, LDA500-M, and LDA1000-M, but the resulting keywords are then mapped to the CSO topics, as for TF-IDF-M.

W2V-W processes the input document with a ten-words sliding window, and uses the word2vec model to identify CSO topics that are semantically similar to the embedding of the window. The embedding of the window are obtained by averaging the embeddings of the single tokens.

STM is the classifier originally adopted by Smart Topic Miner [119], 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 [141]. 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

Table 2.5: Values of precision, recall, and f-measure. In bold the best results.

Classifier	Description	Prec.	Rec.	F1
TF-IDF	TF-IDF	16.7%	24.0%	19.7%
TF-IDF-M	TF-IDF mapped to CSO concepts	40.4%	24.1%	30.1%
LDA100	LDA with 100 topics	5.9%	11.9%	7.9%
LDA500	LDA with 500 topics	4.2%	12.5%	6.3%
LDA1000	LDA with 1,000 topics	3.8%	5.0%	4.3%
LDA100-M	LDA with 100 topics mapped to CSO	9.4%	19.3%	12.6%
LDA500-M	LDA with 500 topics mapped to CSO	9.6%	21.2%	13.2%
LDA1000-M	LDA with 1,000 topics mapped to CSO	12.0%	11.5%	11.7%
W2V-W	W2V on windows of words	41.2%	16.7%	23.8%
STM	Classifier used by STM	**80.8%**	58.2%	67.6%
SYN	Syntactic module	78.3%	63.8%	70.3%
SEM	Semantic module	70.8%	72.2%	71.5%
INT	Intersection of SYN and SEM	79.3%	59.1%	67.7%
CSO-C	The CSO Classifier	73.0%	**75.3%**	**74.1%**

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 data produced in the evaluation, the Python implementation of the approaches, and the word embeddings are available at `http://w3id.org/cso/cso-classifier`. To note that TF-IDF-M, LDA100-M, LDA500-M, LDA1000-M, W2V-W, STM, SYN, SEM, INT, and CSO-C are all general algorithms that classify a text according to the categories from an input taxonomy. Therefore, no method is specifically biased towards CSO.

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 were able to find topics explicitly mentioned in the text, which tend to be very relevant. However, they suffered from a low recall, 58.2%, and 63.8% respectively, as they failed to identify more subtle topics. SEM had lower precision than SYN but higher recall and f-measure, suggesting that it can 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%).

It should be noted that a F1 in the 70%-75% range is remarkably good, given the granularity of the topics in the benchmark, and consistent with the results of other studies that used large classification schemas (e.g., MeSH [50]).

Indeed, the agreement (computed with Fleiss' Kappa) among the three annotators which created the gold standard was 0.451 ± 0.177, indicating a moderate inter-rater agreement [82]. When adding the CSO Classifier as fourth annotator the agreement lowers only slightly to 0.392 ± 0.144. The difference with human annotators may completely disappear when considering a simpler classification schema. A recent experiment using the CSO Classifier for assisting systematic reviews [118] reported that its performance were not statistically significantly different from the ones of six senior researchers (p=0.77) when classifying 25 papers according to five main sub-topics of Software Architecture. We report in Table 2.6 the degree of agreement between the annotator (including also CSO-C), computed as the ratio of papers which were tagged with the same category by both annotators.

Table 2.6: Agreement between annotators (including the CSO classifier) and average agreement of each annotator according to the evaluation in [118]. In bold the best agreements for each annotator.

	CSO-C	User1	User2	User3	User4	User5	User6
CSO-C	-	56%	68%	64%	64%	**76%**	64%
User1	56%	-	40%	**56%**	36%	48%	44%
User2	68%	40%	-	64%	52%	**76%**	64%
User3	64%	56%	64%	-	52%	64%	**68%**
User4	**64%**	36%	52%	52%	-	**64%**	52%
User5	**76%**	48%	76%	64%	64%	-	72%
User6	64%	44%	64%	68%	52%	**72%**	-
Av. Agreement	**66%**	45%	58%	59%	51%	63%	60%

Since its introduction, in 2019, the CSO Classifier was adopted by several applications and research efforts [55, 45, 75, 166]. For instance, [55] used it for annotating the articles from the DBLP computer science library. [45] integrated it in ArtSim, an approach for predicting the popularity of new research papers. [166] classified 1.5M papers and use such topical representation for identifying experts that share similar publishing habits. Finally, [75] developed an ontology-based framework that integrates CSO and the CSO Classifier for retrieving journal articles from academic repositories and dynamically expanding the ontology with new research areas.

2.7.2 Academia/Industry and Industrial Sector Classifications

In order to evaluate the quality of the academia/industry classification in AIDA we randomly 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 authors are reported with affiliation in the industry only;

(iii) 34 collaborative papers, meaning that each paper in this set includes authors with affiliations from academia and authors with affiliations from the industry.

We then asked three independent researchers to manually annotate each paper as ‘academic’, ‘industrial’, or ‘collaborative’ according to the classification above. They were allowed to check online whether a certain institution was academic or industrial. The average agreement score of the three experts was 92.6%. We generated a gold

Table 2.7: Performance of industrial sector classification task.

Industrial Sector	*Precision*	*Recall*	*F1-Score*
Automotive	1.000	1.000	1.000
Healthcare	0.894	0.894	0.894
Computing and it	0.850	0.809	0.829
Electronic	0.700	0.777	0.736
Telecommunication	0.944	0.894	0.918
Macro Average	0.877	0.875	0.875
Weighted Average	0.879	0.875	0.877

standard by using a majority voting strategy. That is, if a paper was considered an 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.

The resulting gold standard perfectly matched the automatic classification.

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 (three senior researchers working within ICT companies and with computer science background) to annotate each organization among the five classes above (or the *other* category if none of the previous categories was appropriate). The average agreement score of the experts was 84.0%.

We 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 performance shown in Table 2.7.

It is interesting to note that, while the performance of our approach is overall quite good, it can differ according to the category. For example it is quite easy to recognize organizations in the 'Automotive' sector, but much less so to identify the ones in 'Electronic'. The same issues also affected human annotators. An analysis of the results seem to suggest that some categories (e.g., Electronic) are potentially more ambiguous according to both human annotators and the linked categories on DBpedia. Conversely, some other categories are more well defined and relatively easy to identify.

In conclusion, the evaluation substantiated that our approaches for classifying documents work remarkably well, performing similarly to human annotators.

2.7.3 Impact Forecasting

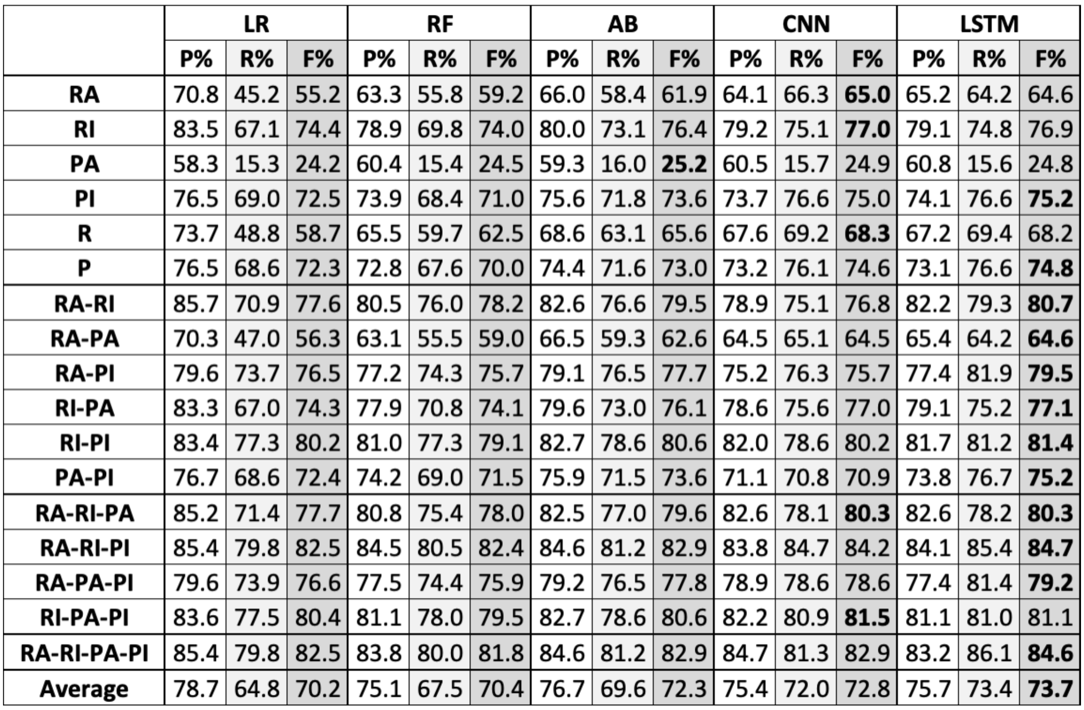
In this section, we present an evaluation of the ability of AIDA to support machine learning forecasters for predicting the impact of research topics on the industry, which is a typical task in the study of academia/industry relationship [8, 48, 176, 94, 127]. Traditionally, the influence of research topics on the industry has been assessed through the quantification of relevant patents. For example, within the AIDA dataset, the research topic labeled as *wearable sensors* was granted a mere 2 patents in the year 2009. Subsequently, in the ensuing years, a multitude of companies increased their investments in this domain, leading to the submission of numerous patents. By 2018, this concerted effort culminated in the issuance of 135 patents in this field. Naturally, the ability to predict such dynamics holds considerable advantages for companies seeking to maintain a position at the forefront of innovation and anticipate evolving market trends.

The literature proposes a range of approaches to patent and technology prediction through patent data, using for instance weighted association rules [8], Bayesian clustering [48], and various statistical models [94] (e.g., Bass, Gompertz, Logistic, and Richards). In the last few years, we saw also the emergence of several approaches based on Neural Networks [176, 127], which lately obtain the most competitive results. However, most of these tools focus only on patents, since they are limited by current datasets that do not typically integrate research articles nor can they distinguish between documents produced by academia or industry. We thus hypothesized that a knowledge graph like AIDA which integrates all the information about publications and patents and their origin should offer a richer set of features, ultimately yielding a better performance in comparison to approaches that rely solely on the number of publications or patents [176, 48, 94, 127].

To examine this hypothesis, we established a gold standard by associating each topic in AIDA with time-frames of five years during which the respective topic had not yet surfaced, as evidenced by fewer than 10 patents. These samples were categorized as *True* if the topic subsequently generated more than 50 industrial patents (PI) over the subsequent 10 years, and classified as *False* otherwise. Each sample was then linked to six time series, encompassing the following: the count of research articles (R), the count of patents (P), the count of research articles from academia (RA), research articles from industry (RI), patents from academia (PA), and patents from industry (PI). As an illustration, consider the sample related to the topic *wearable sensors* in the time-frame 2005-2009. This sample involves six series (R, P, RA, RI, PA, PI), detailing the number of documents within each category during those five years, and it is labeled as *True* because *wearable sensors* eventually produced more than 50 industrial patents (PI) in the subsequent years. In total, the resulting dataset comprises 9,776 labeled samples.

We trained five machine learning classifiers on this gold standard: Logistic Regression (LR), Random Forest (RF), AdaBoost (AB), Convoluted Neural Network (CNN), and Long Short-term Memory Neural Network (LSTM). LR, RF, and

Table 2.8: Performance of the five classifiers on 17 combinations of time series. In bold the best F1 (F) for each combination. The table and the experiments were previously reported in [136].



	LR			RF			AB			CNN			LSTM		
	P%	**R%**	**F%**	**P%**	**R%**	**F%**	**P%**	**R%**	**F%**	**P%**	**R%**	**F%**	**P%**	**R%**	**F%**
RA	70.8	45.2	55.2	63.3	55.8	59.2	66.0	58.4	61.9	64.1	66.3	**65.0**	65.2	64.2	64.6
RI	83.5	67.1	74.4	78.9	69.8	74.0	80.0	73.1	76.4	79.2	75.1	**77.0**	79.1	74.8	76.9
PA	58.3	15.3	24.2	60.4	15.4	24.5	59.3	16.0	**25.2**	60.5	15.7	24.9	60.8	15.6	24.8
PI	76.5	69.0	72.5	73.9	68.4	71.0	75.6	71.8	73.6	73.7	76.6	75.0	74.1	76.6	**75.2**
R	73.7	48.8	58.7	65.5	59.7	62.5	68.6	63.1	65.6	67.6	69.2	**68.3**	67.2	69.4	68.2
P	76.5	68.6	72.3	72.8	67.6	70.0	74.4	71.6	73.0	73.2	76.1	74.6	73.1	76.6	**74.8**
RA-RI	85.7	70.9	77.6	80.5	76.0	78.2	82.6	76.6	79.5	78.9	75.1	76.8	82.2	79.3	**80.7**
RA-PA	70.3	47.0	56.3	63.1	55.5	59.0	66.5	59.3	62.6	64.5	65.1	64.5	65.4	64.2	**64.6**
RA-PI	79.6	73.7	76.5	77.2	74.3	75.7	79.1	76.5	77.7	75.2	76.3	75.7	77.4	81.9	**79.5**
RI-PA	83.3	67.0	74.3	77.9	70.8	74.1	79.6	73.0	76.1	78.6	75.6	77.0	79.1	75.2	**77.1**
RI-PI	83.4	77.3	80.2	81.0	77.3	79.1	82.7	78.6	80.6	82.0	78.6	80.2	81.7	81.2	**81.4**
PA-PI	76.7	68.6	72.4	74.2	69.0	71.5	75.9	71.5	73.6	71.1	70.8	70.9	73.8	76.7	**75.2**
RA-RI-PA	85.2	71.4	77.7	80.8	75.4	78.0	82.5	77.0	79.6	82.6	78.1	**80.3**	82.6	78.2	**80.3**
RA-RI-PI	85.4	79.8	82.5	84.5	80.5	82.4	84.6	81.2	82.9	83.8	84.7	84.2	84.1	85.4	**84.7**
RA-PA-PI	79.6	73.9	76.6	77.5	74.4	75.9	79.2	76.5	77.8	78.9	78.6	78.6	77.4	81.4	**79.2**
RI-PA-PI	83.6	77.5	80.4	81.1	78.0	79.5	82.7	78.6	80.6	82.2	80.9	**81.5**	81.1	81.0	81.1
RA-RI-PA-PI	85.4	79.8	82.5	83.8	80.0	81.8	84.6	81.2	82.9	84.7	81.3	82.9	83.2	86.1	**84.6**
Average	78.7	64.8	70.2	75.1	67.5	70.4	76.7	69.6	72.3	75.4	72.0	72.8	75.7	73.4	**73.7**

AB use the standard implementation of scikit-learn 0.22. CNN and LSTM were implemented using Tensorflow and Keras. CNN was composed of two Convolution1D/MaxPooling1D layers and one output layer computing the softmax function. LSTM uses one LSTM hidden layer of 128 units and one output layer computing the softmax function. We used both binary cross-entropy as loss functions and trained them over 50 epochs.

We ran each of the classifiers on research papers (R), patents (P), and the 15 possible combinations of the other four-time series (RA, RI, PA, PI) to assess which set of features would yield the best results. We performed 10-fold cross-validation of the data and measured the performance of the classifiers by computing the average precision (P), recall (R), and F1 (F). The dataset, the results of experiments, the parameter, and implementation details, and the best models are available at `http://aida.kmi.open.ac.uk`.

Table 2.8 shows the results of our experiment. LSTM outperforms all the other solutions, yielding the highest F1 for 12 of the 17 feature combinations and the highest average F1 (73.7%). CNN (72.8%) and AB (72.3%) also produce competitive results.

As our hypothesis pointed, the utilization of the complete set of features available in AIDA (RA-RI-PA-PI) demonstrates a substantial increase in performance, with a statistically significant improvement ($p < 0.0001$) leading to an F1 score of 84.6%. In contrast, the version employing only the number of patents issued by companies (74.8%) lags behind in performance. Furthermore, when considering the origin of the publications and patents (academia and industry), performance is further enhanced. The inclusion of RA-RI (80.7%) shows a substantial, statistically significant ($p < 0.0001$) improvement over using solely R (68.2%). Similarly, PA-PI (75.2%) exhibits a marginal superiority compared to P (74.8%). These results confirm the value of AIDA's more detailed representation of the document origin in significantly enhancing the predictive performance.

Another interesting outcome is that, when considering only one of the time series, the number of publications from industry (RI) is a significant (p=0.004) better indicator than patents from industry (PI), yielding an F1 of 76.9%, followed by RA, and PA. The best combination of two-time series is RI-PI (81.4%), while the best combination of three-time series is RA-RI-PI (84.7%).

In conclusion, the experiments substantiate the hypothesis that the granular representation of publications and patents in AIDA can support effectively deep learning approaches for forecasting the impact of research topics on the industrial sector. It also validates the intuition that including features from research articles can be very useful when predicting industrial trends.

Chapter 3

AIDA Dashboard

3.1 Introduction

Scientific venues are essential for developing active research communities, promoting the cross-pollination of ideas and technologies, bridging between academia and industry, and disseminating new findings. This is particularly true in the fast-paced field of Computer Science, where conferences are usually the first venue in which researchers present new research efforts [59]. Indeed, each research area in Computer Science is typically associated with a set of venues that help to define and evolve the main challenges and paradigms. Analyzing and monitoring scientific venues is thus crucial for all users who need to take informed decision in this space, such as researchers, scientific editors, developers, government, funding bodies, and other relevant stakeholders.

In this chapter we will analyse the AIDA Dashboard, a web application for analyzing and comparing scientific venues, combining machine learning solutions, semantic technologies, and visual analytics.

The AIDA Dashboard introduces three novel features in order to address the limitations of current tools. First, it provides an interface for comparing and ranking venues within specific fields (e.g., Digital Libraries) according to different metrics and time-frames (e.g., the last five years).

Second, it characterises venues according to 14K research topics from the Computer Science Ontology (CSO). The reader notes that the CSO allows us to structure the research topics within the venues according to a very granular representation [126]. For instance, the topic “Machine Learning” is composed of 760 more specific sub-topics, such as “Denoising Autoencoders” and “Fuzzy Neural Networks”. This allows us to both offer a high-level representation that can be understood by less expert users, but also zoom in on very specific concepts and analyse their trends in time.

Finally, it enables users to analyse the involvement of industry in a venue by i) assessing the impact of commercial organizations across time, ii) reporting the ratio of publications from industry, academia, or collaborative efforts, and iii) categorising industrial contributions according to 66 industrial sectors (e.g., automotive, financial, energy, electronics) from the Industrial Sectors Ontology (INDUSO)[1].

The AIDA Dashboard is available at `http://w3id.org/aida/dashboard`. The current version covers from 1990 to 2022. We are currently working on integrating up-to-date data.

The rest of the chapter is organized as follows. In paragraph 3.2, we review the literature on systems and datasets for assessing scientific venues. Paragraph 3.3, we describe the current limitations of the available tools for analyzing scientific venues. In paragraph 3.4, we describe the AIDA Dashboard in details. Paragraph 3.5 presents the user study.

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

3.2 Background

In this section, we review the relevant literature focusing on two aspects:

i) tools for supporting the assessment of scientific venues, and

ii) scientometrics tools for assessing research trends.

3.2.1 Tools for assessing venues

Several academic search engines and bibliometric tools allow users to explore the venue space. Microsoft Academic, which builds on MAG, provides several analytics about venues. These include number of papers, citations, related conferences, main topics, publications, authors, and main institutions. However, it does not allow users to compare conferences or to analyse the evolution of research topics in time. AMiner and Semantic Scholar allow users to browse venues, but they report only the most prominent authors and the relevant papers. Scholia[2] [110] is a Web service that creates scholarly profiles for topics, people, organizations, and venues according to the information in Wikidata[3]. When a venue is selected, Scholia reports all relevant proceedings, the main articles ranked by their citations, and the main topics, authors, and organizations. However, the data in Wikidata is far from being comprehensive. Moreover, the topics are associated with the venue series as a whole and thus they cannot be used to assess the evolution of the venue across time. The Scopus web application offers several analytics regarding researchers and articles. It links papers to venues, but does not aggregate the latter in venue items. Therefore, it is unable to support significant analyses on venues. Lens.org[4] [72] is a web application that integrates data from MAG, Crossref, Core, and PubMed. It supports the analysis of several scholarly entities such as authors, institutions, contries, journal, conferences, topics, and others. Being based on MAG, it offers the same advantages and limitations of Microsoft Academic.

Overall, all these systems are limited by background data that offer only a coarse-grained representation of venues and their relevant actors (e.g., authors, organizations, countries). For this reason, our first step in the creation of the AIDA Dashboard was the integration and enrichment of several knowledge graphs with the aim of creating more comprehensive metadata about scientific venues.

Our aim, is to identify the main venues in specific fields (e.g., Neural Networks or Digital Libraries instead of the general ones, like Artificial Intelligence), and analyse how they rank in terms of number of publications or average citations as well as whether their scope has changed over the years. To this end, given a venue, we determine its research topics and how they develop over time, so as to understand its status and support stakeholders in making data-informed decisions.

[2] Scholia - https://scholia.toolforge.org

[3] Wikidata - https://www.wikidata.org

[4] Lens.org - https://www.lens.org

3.2.2 Other Scientometric tools

In this section, we report additional state-of-the-art tools, which do not directly support the assessment of venues but have the potential to be extended towards such a direction [163, 64, 158, 77, 5].

Van Eck et al. [163] developed VOSviewer, a tool for creating and visualising networks of publications, researchers, organizations, countries, keywords, and journals. VOSviewer takes as input bibliographic database files (e.g., from Dimensions or Scopus) and builds co-authorship, co-occurrence, citation, bibliographic coupling, or co-citation networks. Ideally, one can download a small dataset concerning a given conference and use such a tool to gain early insights on that conference.

Guilarte et al. [64] developed an interactive tool that leverages citations to visualise branches of science and identify main experts. Specifically, this tool has been applied to the problem of finding potential experts that act as peer reviewers of a target paper. This approach is based on the premise that if a target paper shares similar scientific issues or concerns with some of its references, then the authors of such references can be considered experts. This approach can be potentially extended to analyse whole conference proceedings, to assess the potential experts of that given conference, and even suggest who can act as a programme committee member.

Tosi et al. [158] developed SciKGraph, an approach that takes advantage of semantic technologies and natural language processing to structure research fields from research papers. Specifically, given a corpus of papers, it identifies their concepts and builds a knowledge graph based on their co-occurrence in papers. Concepts are then clustered to show how a scientific area is organised. This approach can be adapted to work on research papers of a single conference to identify its main areas and sub-areas, or analyse research papers of several conferences and identify the similar ones through their topical characterisation.

In general, although these approaches mainly focus on tasks that are different from analysing conferences, with a little adaptation they can support users in improving their understanding of conferences. On the other hand, the AIDA Dashboard focuses specifically on conferences and offers a more integrated suite of analytics in this space.

Furthermore, the above systems do not take into account how much a venue attracts industrial organizations or what relevant industrial sectors are attending the venue. Another goal of the dashboard is to analyse the involvement of the industrial sectors within venues and research topics to provide useful information also to funding agencies.

3.3 Open Issues

Current scholarly search engines and bibliometric applications provide a wide variety of functionalities to support the exploration of research data and produce various kinds of analytics. These include Semantic Scholar[5], Dimensions[6] Scopus[7], Web of Science[8], AMiner[9], and many others. However, these tools only provide a limited set of analytics and metrics for assessing research venues, limiting our ability to perform a comprehensive analysis of these events.

In this paragraph, we focus on three main limitations of these systems. *First*, they do not support a granular comparison of all the venues in a field according to various metrics in time. Google Scholar allows users to rank a limited set of venues, but only according to a course-grained taxonomy of fields and one metric (h5-index). For instance, the field of Artificial intelligence is one of the leaf categories and includes only 20 conferences. Conversely, we would like to identify the main venues in more specific fields, such as Neural Networks or Digital Libraries, how they rank in terms of average citations or number of publications, and how they evolved in the last few years.

Second, current tools do not allow users to analyze the research topics of a venue and their evolution over the years. Conversely, it can be argued that examining these trends is critical to assess the status of a venue and to predict its future performance.

Third, current systems do not take in consideration the industrial involvement in a venue. In particular, they do not report to which degree a venue attracts commercial organizations or what are the relevant industrial sectors. This is a significant missed opportunity since venues are one of the premium public venues in which industry and academia interact and their analysis can offer important insights on how the research in a field is being carried out, supported, or reused by specific industrial sectors. For instance, large tech companies such as Alphabet (Google's parent company), Facebook (now Meta), Microsoft, and IBM became extremely active producing fundamental approaches in the field of Neural Networks in the last few years [99]. Also worth to note that reporting collaborations with non-academic partners is becoming an important metric for funding agencies. Knowledge institutions have to report those to both their funding agencies and the EU. This creates an incentive for academics to collaborate with the industry and to look for suitable venues.

In order to address these issues, we developed the *AIDA Dashboard*, a web application for analyzing and comparing scientific venues which combines machine learning solutions, semantic technologies, and visual analytics. The AIDA Dashboard was developed in collaboration with Springer Nature with the aim of assisting ed-

[5] Semantic Scholar - `https://www.semanticscholar.org/`

[6] Dimensions - `https://app.dimensions.ai/discover/publication`

[7] Scopus - `https://www.scopus.com`

[8] Web of Science - `https://www.webofknowledge.com`

[9] AMiner - `https://www.aminer.cn/`

itors in assessing venues for informing editorial and business decisions. However, it evolved in a more general tool that can produce a wide range of analytics and support multiple use cases.

3.4 Dashboard

In this paragraph we will analyze the process of manipulation of AIDA-KG data in order to: 1) build the analytics for the venues, 2) classification of venues with their research areas of interest, 3) the web interface and its functionalities

3.4.1 The Back-end: Generation of the AIDA Dashboard Dataset

The back-end of the AIDA Dashboard iterates on the venues in AIDA, for each of them computing a set of analytics, and storing the outcome in a collection of JSON files. All the information about a specific conference is thus contained in a single file identified by the conference ID in AIDA. We plan to perform this computation every two months. We label the resulting dataset *The AIDA Dashboard Dataset* and release it to the wide community. The aim is to support other tools as well as further scientometrics analysis. The current version is available at `http://aida.kmi.open.ac.uk/downloads`. We plan to release regular updates of this dataset, every six months.

The AIDA Dashboard dataset describes a venue according to: 1) a set of general metrics, 2) the top authors, organizations, countries, and topics associated with different metrics in time, 3) information about the dynamics between academia and industry in the conference, and 4) the focus areas of the conference. The *focus areas* are a set of high-level topics that the AIDA Dashboard uses for comparing similar conferences. In the following we will detail the process for generating these data from the AIDA knowledge graph.

First, given an input venue, we query the AIDA knowledge graph to gather information such as the name of the venue, its acronym, when it was held, and the total number of publications and citations received by the articles published in its proceedings over the years.

The latter are used to compute h-index, h5-index and the impact factor (over the last 2 years). We compute all these metrics considering the set of papers accepted by the conference, following the same procedure of other systems in this space such as Google Metrics. For instance, we calculate the h5-index over the set of articles published in the conference during the last 5 years.

We then count the number of publications and citations associated with four categories of scholarly items: authors, organizations, countries, and topics. Next, we select the top 100 of each category in terms of publications and the top 100 in terms of citations. Each of the resulting item is associated with its number of publications and citations across the years. For some categories (e.g., authors, organizations) their h-index and h5-index were also computed. Since the distribution of the main topics tends to include several generic high-level topics even when they are under-represented in the specific venue, we also extract an additional set labelled *fingerprint topics*. These are the top 100 topics that in the conference received a percentage of publications and venue higher than their average in the whole Computer Science domain. They are selected by computing the difference between the distribution of topics in the conference and the distribution of the same topics in the whole computer science domain. For instance, the topic *machine learning* is assigned 40% for NeurIPS (Neural Information Processing Systems) because in this conference it appears in about 60% of the articles, while it appears in 20% of the papers in Computer Science.

We then compute the number of publications and citations received from the research papers written by academia, industry, and collaborations, and by the most active industrial sectors.

Finally, we associated the input venue with its main focus areas. Each venue receives a rank in each of these areas based on their average citations in a time interval. For instance, NeurIPS was associated with the focus areas: Neural Networks (2nd overall in the last five years), Machine Learning (2nd), and Artificial Intelligence (5th). The rank allows the users to easily determine the importance of a conference in a field.

In the next paragraph we will describe the algorithm to generate focus areas of a given venue.

Focus areas generation

Algorithm 1 shows the pseudo-code for identifying the focus areas of an input venue. The main purpose of this approach is to determine the research topic that is the most representative of the venue and then returns it together with its super-topics. Simply selecting the topic with the highest frequency is not a good solution since high-level topics are associated with all the publications of their sub-topics. For instance, a naïve algorithm based on frequency may assign to NeurIPS the focus area *artificial intelligence*, ignoring what component of AI is more prominent in this case. Conversely, we may detect that the large number of publications associated with AI is mainly due to the prominence of the sub-topic *machine learning*, and in turn that the majority of articles associated with this area are from the specific sub-topic *neural networks*. Therefore, our approach first orders the topics according to their number of publications (line 1). Topics are then (line 2) filtered by using a whitelist. Next, we fetch (line 3) the total number of publications of the venue. The algorithm iterates on all the topics (line 6) and selects the first topic as candidate focus area (lines 9-11). For the other topics, it checks whether it is a descendant of the current candidate (first condition, line 12), and if it is the main reason for its high frequency of publications (second condition, line 12). It does so by assessing

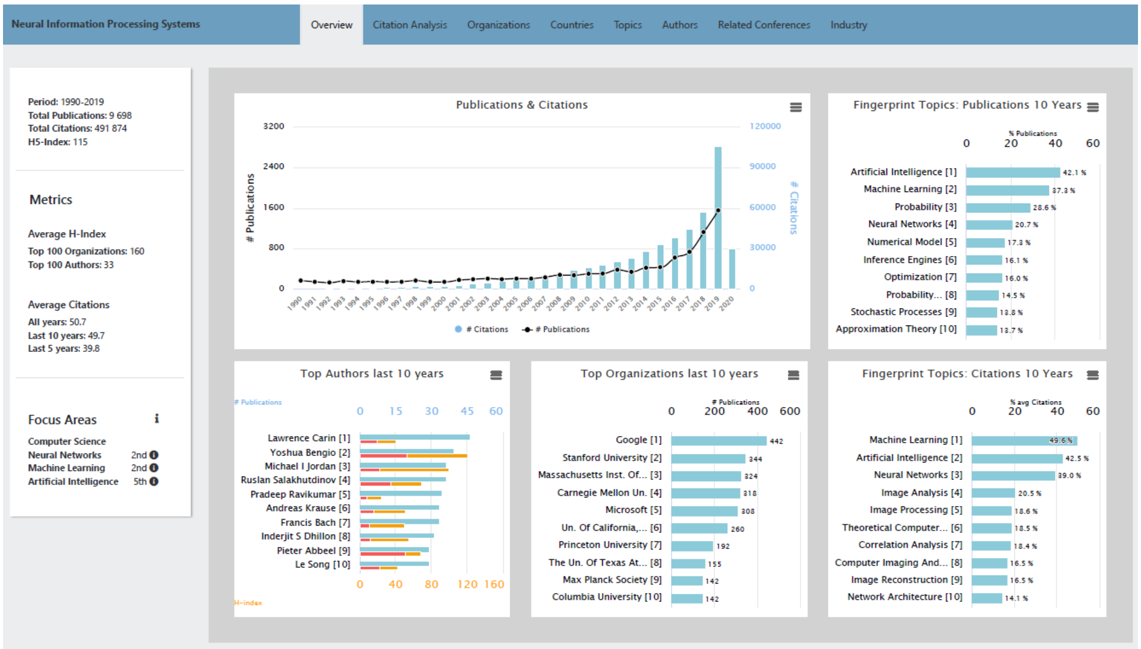
Algorithm 1: Focus Areas Generation.

```
Input  : Venue ID venue, Threshold for taking a sub-area subtopic_thr,
         Whitelist of areas whitelist
Output: Set of Focus Areas focus_areas
1  topics ← getVenueSortedTopics (venue);
2  topics ← filter (whitelist, topics);
3  publications_c ← getTotalPublications (venue);
4  candidate ← NULL;
5  candidate_impact ← 0;
6  foreach topic in topics do
7      publications_t ← getNumberOfPubs (topic, venue);
8      impact = publications_t/publications_c;
9      if candidate is NULL then
10         candidate ← topic;
11         candidate_impact ← impact;
12     else if (topic is descendant of the candidate) AND
        (impact/candidate_impact > subtopic_thr) then
13         candidate ← topic;
14         candidate_impact ← impact;
15 focus_areas ← expand (candidate);
16 return focus_areas
```

if the percentage of the candidate publications associated also with the sub-topics is higher than a threshold (line 12, *subtopic_thr*=0.6 in the prototype). If this is the case, it selects the sub-topic as new candidate (lines 13-14). Finally, it returns (lines 15-16) the last candidate topic (e.g., neural networks) and all its super topics (e.g., machine learning, artificial intelligence). When computing the focus areas for all venues in Computer Science, the whitelist was first initialised to the full set of topics in CSO. We then analyzed the distribution of the resulting focus areas and generated a whitelist including the 166 focus areas that were associated with at least 5 conferences. The purpose of this operation is to discard minor areas not useful for comparing a fair number of conferences and obtain a representative whitelist which we feed to sequent executions of the algorithm. This whole process takes a few minutes on an average machine and it is processed offline once a year.

3.4.2 The Web Interface

The Web interface of the AIDA Dashboard allows users to search for the full name or the acronym of a conference using an autocomplete field. When a venue is selected, it loads the corresponding JSON file from the back-end. It then produces interactive views of the resulting analytics structured in eight tabs: *Overview*, *Citation Analysis*, *organizations*, *Countries*, *Authors*, *Topics*, *Related Venues*, and *Industry*.


Neural Information Processing Systems
Overview
Citation Analysis
Organizations
Countries
Topics
Authors
Related Conferences
Industry
Publications & Citations
3200
2400
1600
800
0
# Publications
120000
90000
60000
30000
0
# Citations
1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020
● # Citations
# Publications
Fingerprint Topics: Publications 10 Years
% Publications
0 20 40 60
Artificial Intelligence [1] 42.1 %
Machine Learning [2] 37.3 %
Probability [3] 28.6 %
Neural Networks [4] 20.7 %
Numerical Model [5] 17.3 %
Inference Engines [6] 16.1 %
Optimization [7] 16.0 %
Probability... [8] 14.5 %
Stochastic Processes [9] 13.8 %
Approximation Theory [10] 13.7 %
Top Authors last 10 years
# Publications
0 15 30 45 60
Lawrence Carin [1]
Yoshua Bengio [2]
Michael I Jordan [3]
Ruslan Salakhutdinov [4]
Pradeep Ravikumar [5]
Andreas Krause [6]
Francis Bach [7]
Inderjit S Dhillon [8]
Pieter Abbeel [9]
Le Song [10]
0 40 80 120 160
H-index
Top Organizations last 10 years
# Publications
0 200 400 600
Google [1] 442
Stanford University [2] 344
Massachusetts Inst. Of... [3] 324
Carnegie Mellon Un. [4] 318
Microsoft [5] 308
Un. Of California,... [6] 260
Princeton University [7] 192
The Un. Of Texas At... [8] 155
Max Planck Society [9] 142
Columbia University [10] 142
Fingerprint Topics: Citations 10 Years
% avg Citations
0 20 40 60
Machine Learning [1] 49.6 %
Artificial Intelligence [2] 42.5 %
Neural Networks [3] 39.0 %
Image Analysis [4] 20.5 %
Image Processing [5] 18.6 %
Theoretical Computer... [6] 18.5 %
Correlation Analysis [7] 18.4 %
Computer Imaging And... [8] 16.5 %
Image Reconstruction [9] 16.5 %
Network Architecture [10] 14.1 %

Period: 1990-2019
Total Publications: 9 698
Total Citations: 491 874
H5-Index: 115

Metrics

Average H-Index
Top 100 Organizations: 160
Top 100 Authors: 33

Average Citations
All years: 50.7
Last 10 years: 49.7
Last 5 years: 39.8

Focus Areas ℹ

Computer Science	
Neural Networks	2nd ℹ
Machine Learning	2nd ℹ
Artificial Intelligence	5th ℹ

Figure 3.1: AIDA Dashboard - the Overview tab of the NeurIPS conference.

The **Overview** tab is the introductory page of a venue, where the user is first redirected. It provides general information about the conference performance and trends. Figure 3.1 shows as example the Overview tab of the NeurIPS conference. This page is organized in two sections. The bar on the left gives information and metrics (e.g., the period of activity, the total number of publications and citations, the h5-index) about the underlying conference. It also provides general information about the average h-index of the organizations and authors who published in the conference as well as the average citations received by the published papers. In the lower part, it reports the focus areas and the rank of the conference in each of them (according to the average citations in the last 5 years). The section on the right provides several charts about the number of publications and citations over the years, the main authors and organizations in terms of publications (in the last 10 years), and the top fingerprint topics in terms of publications and citations (in the last 10 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 venue in the focus areas. For instance, in Figure 3.2 we can see that NeurIPS has been among the top two conferences in Neural Networks and Machine Learning and the top ten conferences in Artificial Intelligence for the last 20 years. This visualization is typically used by Springer Nature editors to assess the performance of venues within different communities.

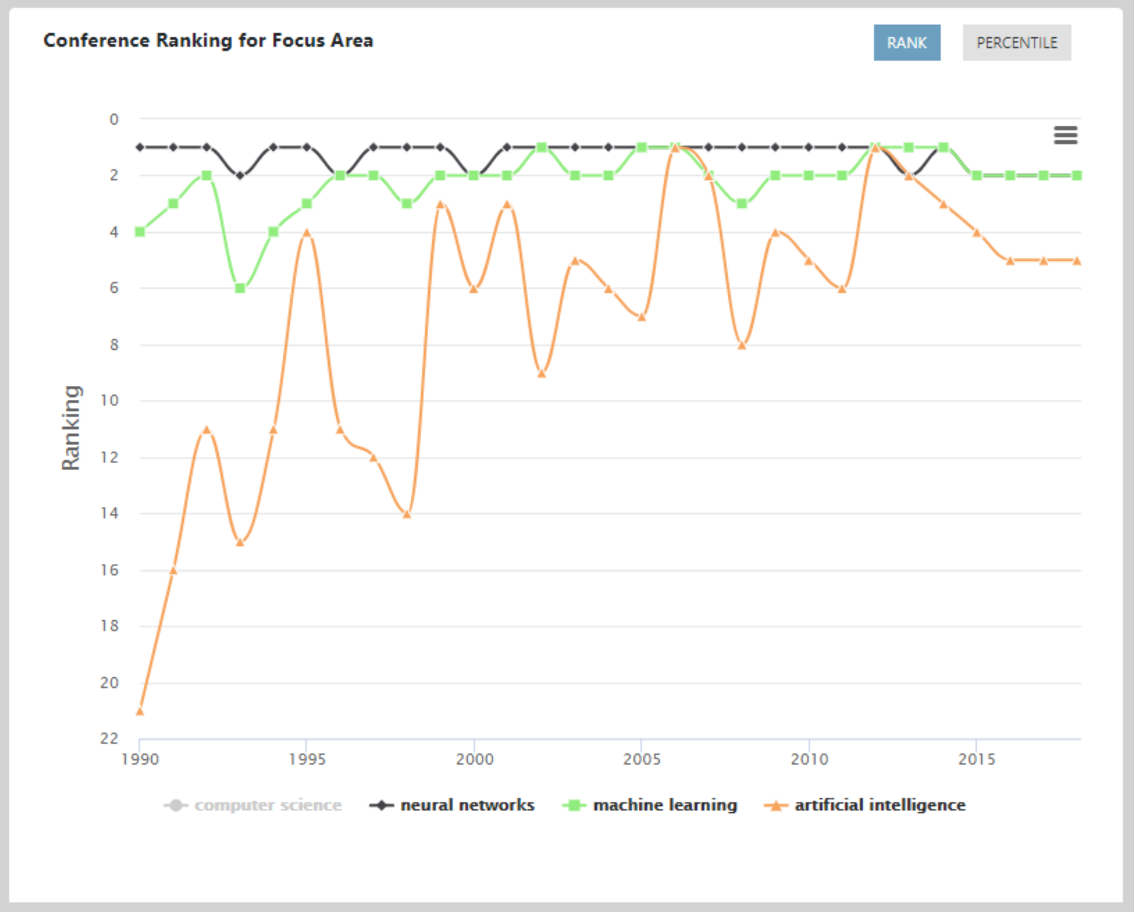
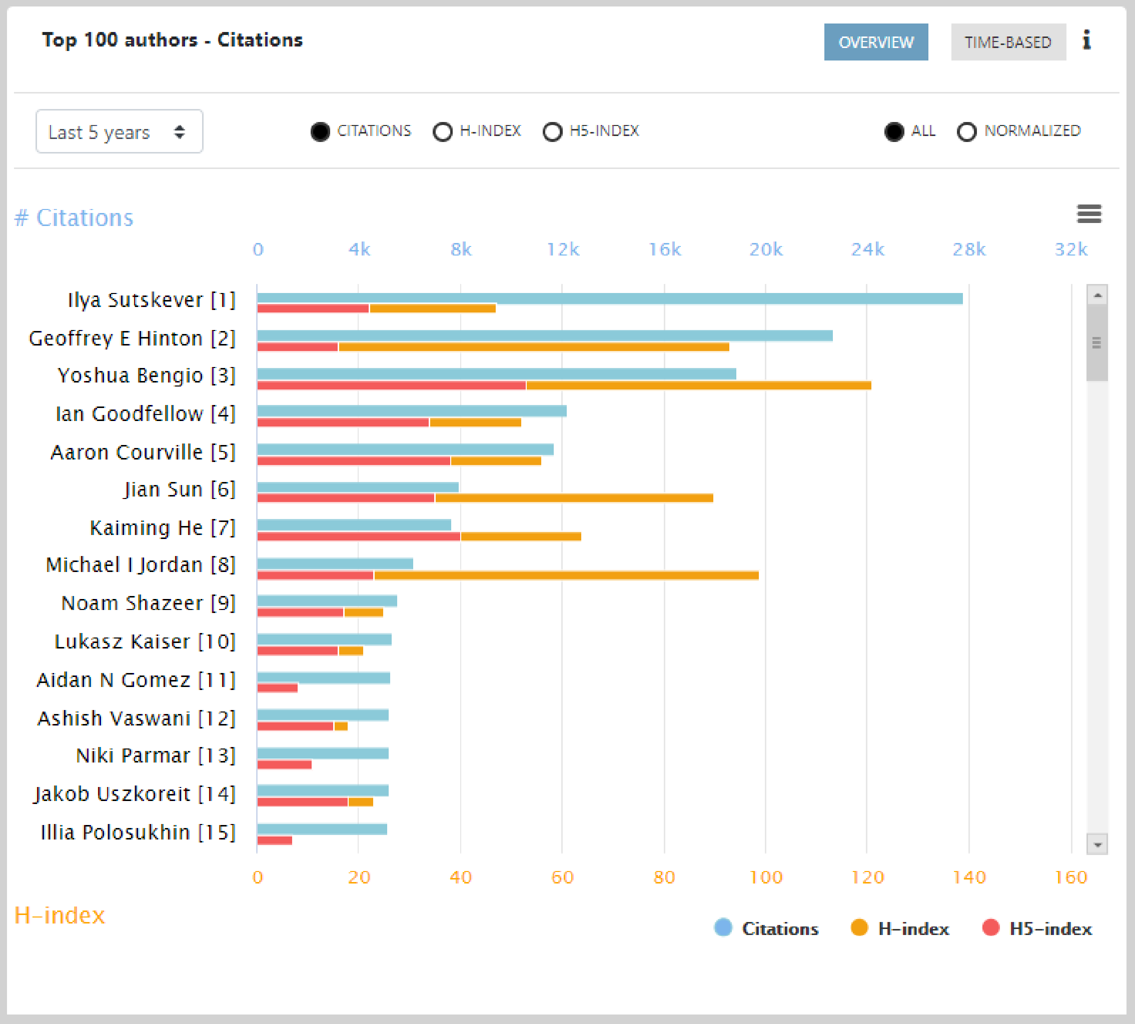


Figure 3.2: Portion of the Citation Analysis tab - The ranking of NeurIPS its focus areas.

The **Organizations** tab shows several analytics about the main institutions active in the venue. In this section the users can assess the main organizations according to their number of publications, citations, and average citation. Organizations can also be filtered according to their types (academia, industry, or all). The default interface used by the dashboard for reporting these data is a bar chart in which each item is associated with the total of the metric in a period (e.g., last five years). The user can also change this view (using the 'time-based' button) to a line-chart showing the same data across the years, which allows users to easily analyze trends in time.

The **Authors** tab uses the same interface for displaying the main researchers associated with their number of publications, citations, and average citations. The researchers can also be sorted by their overall H-index and H5-index, in order to quickly identify high impact researchers. Figure 3.3 shows the authors from NeurISP ordered according to their number of citations in the last five years. Editors at Springer Nature typically use the Organizations and Authors tabs to assess the



Top 100 authors - Citations

OVERVIEW TIME-BASED

Last 5 years

● CITATIONS ○ H-INDEX ○ H5-INDEX

● ALL ○ NORMALIZED

Figure 3.3: Portion of the Authors tab - Authors ranked by citations in NeurIPS.

quality of researchers and organizations attracted by the venues. This is particularly important for assessing relatively young venues that may not yet have developed a strong citation record.

The **Countries** tab allows the users to analyze the contribution of specific countries. The user can switch between the Chart view and the Map view. The first one shows the set of countries according to their number of publications, citations, and average citations. The second view arranges the information about the frequency of articles by country in a world map.

The **Topic** tab allows the users to analyze the topic trends over time. Specifically, it shows two selections of topics: main topics and fingerprint topics, discussed earlier in the paper. Figure 3.4 shows the main topics of NeurIPS. On the left side we indicate the percentage of publications in which the underlying topic appears. On the right side we show the number of citations received by articles in which the topic

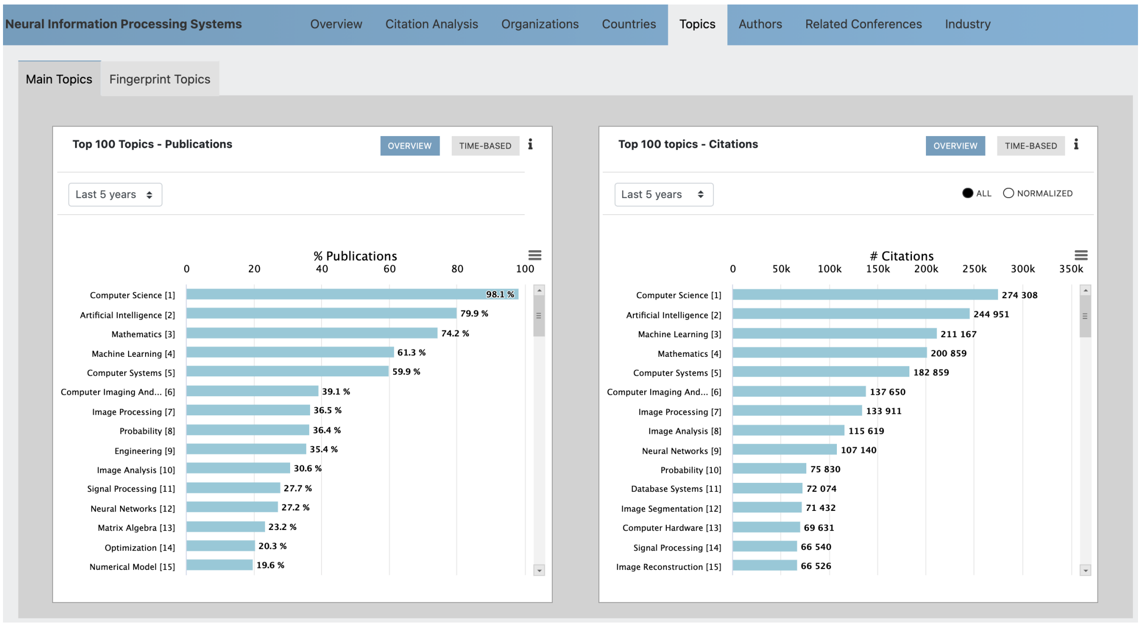

Neural Information Processing Systems
Overview
Citation Analysis
Organizations
Countries
Topics
Authors
Related Conferences
Industry
Main Topics
Fingerprint Topics
Top 100 Topics - Publications
OVERVIEW
TIME-BASED
Last 5 years
% Publications
0
20
40
60
80
100
Computer Science [1] 98.1 %
Artificial Intelligence [2] 79.9 %
Mathematics [3] 74.2 %
Machine Learning [4] 61.3 %
Computer Systems [5] 59.9 %
Computer Imaging And... [6] 39.1 %
Image Processing [7] 36.5 %
Probability [8] 36.4 %
Engineering [9] 35.4 %
Image Analysis [10] 30.6 %
Signal Processing [11] 27.7 %
Neural Networks [12] 27.2 %
Matrix Algebra [13] 23.2 %
Optimization [14] 20.3 %
Numerical Model [15] 19.6 %
Top 100 topics - Citations
OVERVIEW
TIME-BASED
Last 5 years
● ALL
○ NORMALIZED
# Citations
0
50k
100k
150k
200k
250k
300k
350k
Computer Science [1] 274 308
Artificial Intelligence [2] 244 951
Machine Learning [3] 211 167
Mathematics [4] 200 859
Computer Systems [5] 182 859
Computer Imaging And... [6] 137 650
Image Processing [7] 133 911
Image Analysis [8] 115 619
Neural Networks [9] 107 140
Probability [10] 75 830
Database Systems [11] 72 074
Image Segmentation [12] 71 432
Computer Hardware [13] 69 631
Signal Processing [14] 66 540
Image Reconstruction [15] 66 526

Figure 3.4: AIDA Dashboard - the Topics tab of the NeurIPS conference.

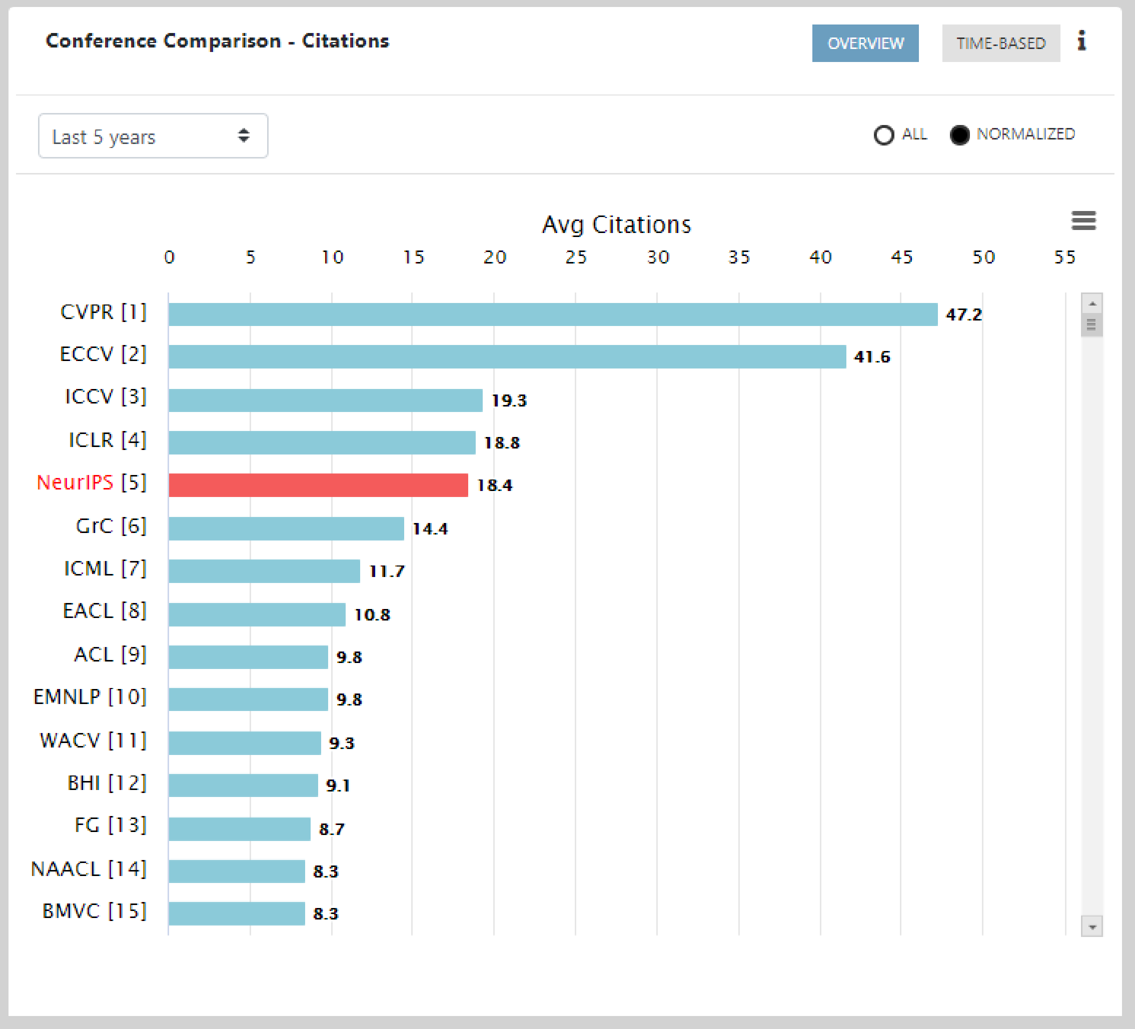
appears.

The **Related Venues** tab allows the users to compare the underlying venue against all the others venues of the same type in the same fields according to their number of publications, citations, and average citations for paper (e.g., if a conference is selected it will be compared against all the other conferences in the same field, if a journal is selected it will be compared against all the other journals). The user can contextualise the comparison to different fields. For example, the NeurIPS conference can be compared with all the other conferences in the fields of Neural Networks, Machine Learning, and Artificial Intelligence. Figure 3.5 shows the comparison of NeurIPS with the other top conferences in Artificial Intelligence. The conference in analysis is highlighted in red.

Finally, the **Industry** tab reports the number 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 industrial sectors. Figure 3.6 shows the trend of publications received by companies in different industrial sectors.

3.4.3 Advanced Search

Figure 3.7 displays the Advanced Search panel, which allows users to browse and compare venues according to their fields. The user can browse the different fields using the selection menus and switch between journals and conferences with the button in the upper right. For instance, a user checking all the conferences in the



Conference Comparison - Citations

Last 5 years

○ ALL ● NORMALIZED

Figure 3.5: Portion of the Related Conferences tab - Conferences in Artificial Intelligence ranked by average citations.

field *"The Web"* can decide to focus further the analysis and only show the subset of venues within the sub-area *"Semantic Web"*. Clicking on a specific venue will bring the user to the relevant venue panel.

Journals and conferences can be ranked according several metrics, including:

a) average citations received in the last five years,

b) average articles published in the last 5 years,

c) h5-index,

d) the average h5-index of the relevant organisations, and

e) the average h5-index of the relevant authors.

The last two metrics are not typically offered by alternative systems, but are very useful to identify emergent conferences that are attracting strong research groups but may not have yet received a good number of citations. Venues can be also ranked according to the set of external ratings discussed in Section 2.2.

3.5 Evaluation

In this paragraph, we discuss the results of a user study involving 10 senior researchers.

3.5.1 User Study

We performed a user study on the AIDA Dashboard to assess the quality and usefulness of the analytics as well as the usability of the user interface. To this end, we organised individual sessions with 5 SN editors and 5 researchers in Computer Science. In each session, we first presented the AIDA Dashboard 2.0 for about 20 minutes. We then assigned to the users the task of analysing two venues and a focus area of their expertise in order to assess the quality of the resulting analytics. After the hands-on session the users filled a five-parts survey about their experience. The first part covered the users background and expertise. The second part was a standard System Usability Scale[10] (SUS) [39] questionnaire to gauge the usability of the AIDA dashboard. The third section asked the users to rate the quality of the analytics for the two venues and the focus area on a Likert scale in the [1-5] range. The fourth part included four open questions about strengths and weaknesses of the dashboard asked to all users and two further questions that were asked only to the editors. Finally, the fifth part asked to list at least three of the most useful functionalities.

The data produced during the user study are available online[11].

User Background

The five researchers in the user study are all senior researchers, with an average of 13.4 years of experience, and come from different institutions:

i) University of Cagliari (IT),

ii) Institute for Applied Informatics (DE),

iii) FIZ Karlsruhe – Leibniz (DE),

iv) University of Paris 13 (FR), and

[10] System Usability Scale (SUS) - `https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html`

[11] AIDA Evaluations - `https://w3id.org/aida/downloads#evaluation`

v) National Council of Research (IT).

The five editors are at various career stages (1, 5, 13, 21, and 25 years of experience) and come from different departments within SN.

The areas of expertise of the 10 users include Artificial Intelligence, Natural Language Processing, Semantic Web, Robotics, Machine Learning, Multimedia Systems, and Theoretical Computer Science.

SUS questionnaire

The SUS questionnaire provided excellent results obtaining a score of 88.5/100 considering all users. This corresponds to the 97% percentile rank in terms of usability (A+ grade) according to the SUS guidelines[12]. In general, editors were more severe than researchers, mostly because they consider the dashboard an important working tool and they where very motivated in suggesting further improvements. Indeed, editors scored an average 84.5 SUS score (96% percentile rank), while researchers yielded 92.5 (98%). This version of the dashboard (2.0) showed a better usability than the previous one, which achieved a SUS score of 87.5 in a user study involving 10 researchers [22].

Figure 3.8 reports the average score given by researchers (red bars) and editors (blue bars) to specific questions in the SUS questionnaire. Odd questions are positive (a higher score is better) while even ones are negative (a lower score is better).

Overall, all the users found the system very easy to use (high values in question 3), they could easily learn the system (question 7), and they do not need support to use the system (question 4). The editors found some inconsistency in the integration of the functionalities (question 5). Finally, all users would like to frequently use the dashboard (question 1).

Quality Assessment

We asked the users to evaluate the quality of the analytics produced by the AIDA Dashboard for the two venues and the focus area according to a Likert scale. On average, editors scored 3.8 for venues and 4 for focus areas, whereas researchers 4.2 for both venues and focus areas.

The range of fields and venues analysed by the users included Artificial Intelligence (AAAI, ICML, EANN, NC&L, Machine Learning), Natural Language Processing (EMNLP, ACL, EACL), Multimedia Systems (ACM Multimedia, Multimedia Tools & Applications), Robotics (ICRA, IROS), The Web (The Web Conference), Information Retrieval (SIGIR), Digital Library (TPDL), Semantic Web (ISWC), and Theoretical Computer Science (Information & Computation, iConference).

[12] Interpreting a SUS score - https://measuringu.com/interpret-sus-score/

Open Questions

We summarise here the main feedback emerged from questions Q1-Q4 (all users) and questions Q5-Q6 (only editors).

Q1. What are the main strengths of AIDA Dashboard? Users were positively impressed by the easy and intuitive interface and the large amount of analytics. Other positive feedback regarded the granularity of the topic classification and the fact that the system addressed a real need in the community, i.e. analysing and comparing venues.

Q2. What are the main weaknesses of AIDA Dashboard? Users listed a range of issues that we plan to address in the future. One researcher suggested that the major limitation is that the coverage is constrained to the Computer Science domain. Another one reported some disambiguation issues, in particular regarding authors with similar names. One more suggested that certain functionalities were hard to locate because the second level tabs were not particularly discernible. One editor mentioned the need of analysing venues in time ranges smaller than 5 years. Another one criticised the current interface for navigating the taxonomy based on selection menu. Finally, one editor did not find smooth the integration of journals and conferences and asked to be able to compare both of them in the same panel.

Q3. Can you think of any additional features to be included in AIDA Dashboard? Researchers mentioned: 1) adding more type of scholarly entities to analyse (e.g., organisations, researchers), 2) the ability to compare specific charts from different venues, 3) some additional metrics (e.g., number of papers that contributed to the citation count), 4) various minor GUI improvements, and 5) the ability to rank topics alphabetically. Editors mentioned:

1) the ability to directly compare conferences to journals; 2) a better integration with the CSO taxonomy; 3) adding information about the publishers of the venues, and 4) considering also books series.

Q4. How comprehensive/accurate do you consider the list of focus areas associated with the venues in AIDA Dashboard? All the researchers found the list of focus areas accurate and comprehensive. However, two of them suggest that they were sometimes too broad and would have liked the ability to browse venues also according to arbitrary research topics. Four editors found the list very accurate and comprehensive, while one of them identified some missing areas in their field of expertise and suggest edits for the Machine Learning branch (already implemented in the current version).

Q5. In which way the AIDA Dashboard support your work? Two editors reported that the system was very useful for supporting junior or new editors in analysing specific research fields. Two found it very helpful in identifying notable trends in venues topics and performing country-centric analysis. One found it very useful in identifying and comparing venues. Some editors also highlighted how the dashboard supports the detection of conferences and workshops that could produce special issues about specific emerging topics.

Q6. What competitive advantages would you say the AIDA Dashboard provides with respect to Scopus/Google Scholar (if any)? One editor pointed out that the AIDA Dashboard provides better visualisations as well as more granular analytics compared to Scopus and Google Scholar. One considered the auto-suggested search more helpful and simpler than the one in Scopus search. Finally, an editor found the AIDA Dashboard more powerful in analysing conferences and journals, preferring instead Google Scholar for analysing individual researchers or articles.

Best Functionalities

We asked the ten users to list at least three of the most useful sections of the AIDA Dashboard. Figure 3.9 reports the user preferences. The *Related Conferences/Journals* tab was the most appreciated section for both editors and researchers. This highlights how comparing venues is a critical task that was not well supported by previous solutions. Interestingly, researchers preferred the analytics about topics and citation analysis, while editors the analysis on authors and organisations.

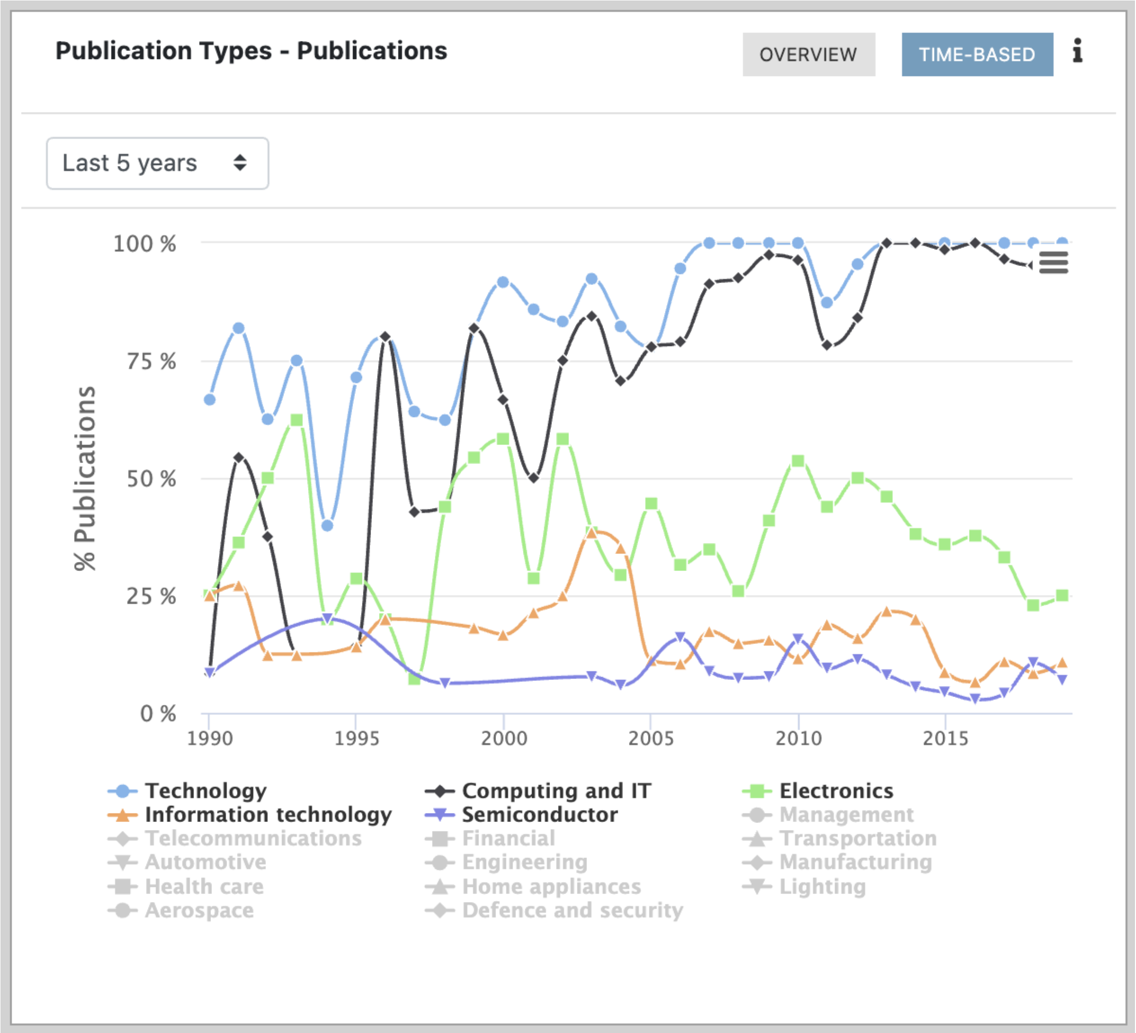
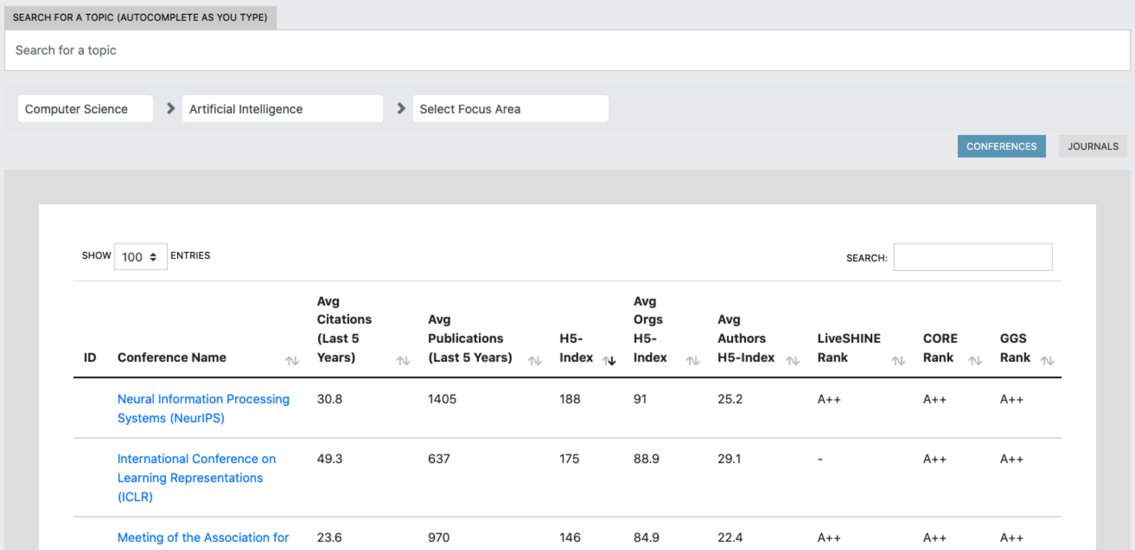


Figure 3.6: Portion of the Industry tab - The main industrial sectors in NeurIPS across time. The percentage indicates the fraction of papers published in the corresponding year by companies of the underlying industrial sector.



SEARCH FOR A TOPIC (AUTOCOMPLETE AS YOU TYPE)

Search for a topic

Computer Science > Artificial Intelligence > Select Focus Area

CONFERENCES JOURNALS

SHOW 100 ENTRIES SEARCH:

ID	Conference Name	Avg Citations (Last 5 Years)	Avg Publications (Last 5 Years)	H5-Index	Avg Orgs H5-Index	Avg Authors H5-Index	LiveSHINE Rank	CORE Rank	GGS Rank
	Neural Information Processing Systems (NeurIPS)	30.8	1405	188	91	25.2	A++	A++	A++
	International Conference on Learning Representations (ICLR)	49.3	637	175	88.9	29.1	-	A++	A++
	Meeting of the Association for	23.6	970	146	84.9	22.4	A++	A++	A++

Figure 3.7: The Advanced Search Panel displaying conferences in Artificial Intelligence ranked by h5-index.

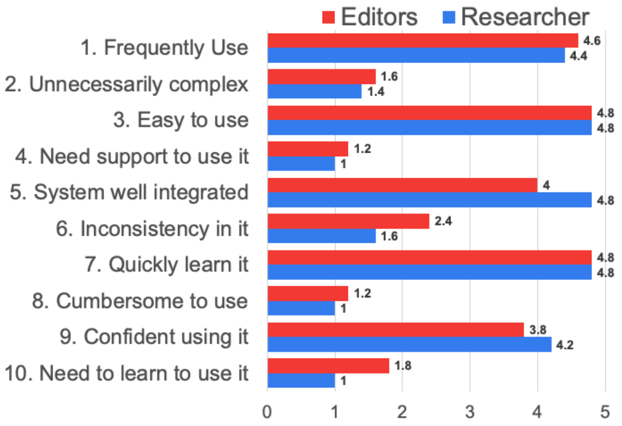


Figure 3.8: The SUS Questionnaire results.

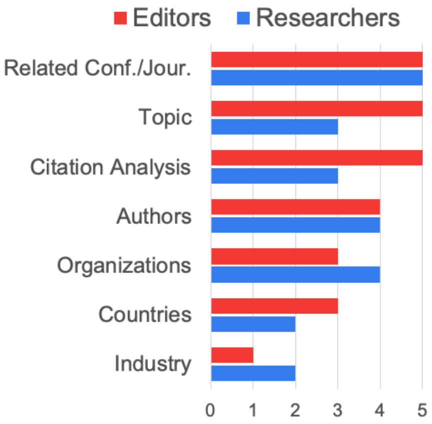


Figure 3.9: Number of votes received by each Section/Functionality.

Chapter 4

AIDA-Bot

4.1 Introduction

In recent years, chatbots have gained widespread acceptance and are now extensively utilized across various domains to streamline and automate communication on a large scale. These chatbots have emerged as vital tools for assisting users in addressing their inquiries and performing a wide array of tasks, including customer support [174, 43], item ordering [51], ticket booking [147], providing driving helps [76], and more. Typically, these chatbots employ advanced techniques like natural language understanding and generation to comprehend user queries. They then construct an equivalent query for a knowledge base and furnish users with information based on the results of the generated query [4].

Conversational tools are now prevalent across various sectors. In e-learning, they introduce advanced communicative features, enhancing the educational process. Their presence has been shown to boost student motivation and engagement, leading to an uptick in meta-cognitive skill acquisition [56].

In the public administrations, these conversational tools have been adopted for diverse purposes. For instance, a contemporary Italian job portal has integrated a chatbot that suggests job opportunities based on user skills [30]. Another initiative presents a chatbot system tailored to address queries about services rendered by public entities [90]. Challenges were related to the big set of services, their complexity, the specific domain of public service, user query phrasing, and the linguistic gap between professionals (like attorneys or bureaucrats) and the general public. A study by Van Noordt and colleagues [164] delves into the exploration of three chatbots used in the public sectors of Vienna, Latvia, and Bonn, concluding that chatbot integration in these settings often accompanies minor organizational shifts.

In healthcare, there's a great focus on leveraging recent AI advancements to streamline services in facilities like nursing homes and hospitals [43]. Callejas and Griol's research [42] sheds light on how conversational platforms are being utilized in mental health care. Additionally, a comprehensive review by Montenegro et al. [101] examines around 4,145 papers discussing health-centric conversational tools from 2009 to 2019. In the aerospace sector, conversational tools have been suggested as a means to provide swift and precise responses in intricate scenarios. A notable example by Liu et al. [88] integrates a task-driven dialogue system with a conversational tool and an interactive Q&A module, aiming to evaluate the advantages of intelligent and conversational searches in cockpit documentation.

Recently, we have witnessed an increasing diffusion and employment of Knowledge Graphs (KGs), which are becoming a standard solution for representing complex interconnected data. KGs acquire and integrate information from the real world by using an ontology. In particular, it represents the information by means of a graph whose nodes are entities whereas edges represent their relationships [67]. This formal and structured representation allows automatic programs to better interpret users' questions.

In this context, the new challenge is to design chatbot architectures able to access and operate together with KGs and extend the range of queries that users are allowed to express in order to identify and return the information they might be interested in. For example, Bockhorst et al. [34] present an approach to developing task-oriented conversational interfaces that construct a system of grammar to correctly infer parses from natural language. The system grammar is built by leveraging the structured types and entities of an underlying KG complemented by a machine learning-driven restructuring procedure. Developing a new generation of chatbots able to capitalize on knowledge graphs is thus the natural but challenging step forward.

However, while there is an abundance of literature about conversational agents in education [170], to the best of our knowledge, works on conversational agents for supporting academic and scientific research are not present within the scientific literature.

In this chapter, we introduce AIDA-Bot, a chatbot able to answer various questions about the research landscape and the scientific literature. This conversational agent has been designed to both 1) support a set of predetermined question types (e.g., “List all entities with a certain characteristic”, “Compare two entities”) by automatically translating them to formal queries on the knowledge graph, and 2) answer open questions (e.g., “What is a convolutional neural network?”, “Define knowledge graph”) by summarising information from relevant articles in the knowledge graph. This hybrid approach ensures that the responses provided are grounded in factual information that can be easily verified and, if necessary, corrected by updating the knowledge graph.

The rest of this chapter is structured as follows. In Section 4.2, we report previous related work on conversational agents . Section 4.3 introduces the architecture of the chatbot. Section 4.4 delineates the outcome of the qualitative evaluation.

4.2 Background

Chatbot technology has always been attractive to researchers for decades. It traces back to 1966 when Joseph Weizen-Baum developed ELIZA[1], an early example of conversational software. ELIZA used keyword matching and context identification to engage with users, although it couldn’t sustain extended conversations. Another notable historical chatbot is ALICE[2], which won the Loebner Prize award three times (in 2000, 2001, and 2004). ALICE is built on the Artificial Intelligence Markup Language (AIML)[3], a lightweight and highly configurable language that still underpins many contemporary chatbots [3].

Practitioners in the field continually work on developing and studying new features to enhance the functionality of existing methods. They have also introduced

[1] https://en.wikipedia.org/wiki/ELIZA

[2] A.L.I.C.E. (Artificial Linguistic Internet Computer Entity) - https://en.wikipedia.org/wiki/Artificial_Linguistic_Internet_Computer_Entity

[3] http://www.aiml.foundation/

new architectural approaches. These advancements often make use of ontologies and context, including information about both the ongoing and previous conversations [128]. Chatbots can be categorized based on various characteristics: i) knowledge domain, ii) type of interaction, iii) usage, iv) design techniques [70]. The last describes the design philosophy behind a chatbot and how different categories of chatbots deal with the conversation in a given context

When considering their objectives, chatbots may be categorized into two main classes: task-oriented [130] and non-task-oriented [46].

Task-oriented chatbots are specialized for particular scenarios, such as booking accommodations, ordering products, or assisting users in obtaining specific information. These chatbots are focused on helping users achieve a particular objective within a defined domain but typically lack general knowledge [85]. In contrast, non-task-oriented chatbots are primarily designed for extended conversations and operate in open domains. They aim to emulate the characteristics of unstructured human-human conversations and are not limited to specific tasks [131].

We can further categorize chatbots based on their mode of interaction, distinguishing between text-based and voice-based chatbots. Text-based chatbots engage users through written messages, with the primary goal of promptly identifying user needs and providing instant solutions. Businesses frequently employ text-based chatbots to manage interactions with their customers [35]. One notable advantage of text-based chatbots is their adaptability for integration with various platforms, including social media and messaging applications. In contrast, voice-based chatbots [134, 6] are capable of recognizing human speech and responding with synthesized vocal responses. Prominent examples include personal assistants like Amazon Alexa, Google Assistant, Apple's Siri, and Microsoft's Cortana. These voice-based chatbots are commonly used for task-oriented purposes, such as web searches, making phone calls, sending text messages, playing multimedia content, interacting with Internet of Things (IoT) devices, and even providing entertainment through jokes [10].

It is also possible to characterise chatbots according to their engine (rule-based vs AI-based). Rule-based chatbots [149] use a tree-like flow to help users with their questions. This means that they guide the user with follow-up questions to eventually get the correct response. The structures and answers are typically predefined. Other chatbots employ AI and natural language processing techniques [100], that, unlike rule-based chatbots, do not use keywords, patterns, or rules to determine the user's intent, but try to infer it directly from the text.

Sometimes, chatbots are tailored to work in specific domains such as i) healthcare [83], ii) education [113], and iii) business [27]. Chatbots in healthcare support patients and their relatives by answering specific health-related questions on HIV/AIDS [38], child health [162], and mental health [112], to name a few [83]. For example, Divya et al. [54] developed a medical chatbot for self-diagnosing diseases, which provides also detailed descriptions of them. Additional chatbots in healthcare include MedChatbot [31] and Mandy [108]. The former is used to support medical students. The latter is used by healthcare workers to automate patient intake. Other chatbots collect information about people's diet [157] or provide restaurants with a tool to collect allergy information based on users' allergens [68].

Chatbots in education support the teaching of a variety of subjects, such as English [143], Medicine [31], and business process models [132]. Some chatbots are also able to answer university-related questions that are typically found in FAQs [129]. The reader is referred to [113] for a review of works on the use of chatbots in education. Finally, in the business domain, there are chatbots supporting companies in their daily tasks [27]. For example, chatbots were developed to support customer service for businesses and e-commerce [174, 51], helping to complete certain tasks [87], and improve user experience [52]. Works presented in a recent workshop [1] discussed innovative techniques to interact with chatbots, understand conversations, promote mental health and well-being, improve the coverage of clarification responses, assess chatbot applications in different domains, and measure how a chatbot can be supportive or engaging.

In the last few years, we saw the emergence of a variety of conversational agents and question-answering systems that build on semantic web technologies and knowledge graphs [66, 25, 153, 2, 120]. The main advantage of these solutions is their ability to integrate and formulate complex queries on heterogeneous data from multiple sources [58], including large-scale general knowledge bases such as Wikidata [102] and DBpedia [25]. They also support reasoning and link prediction techniques [103] for identifying and correcting errors as well as enriching the knowledge base with new facts [121]. This allows a conversational agent to act accordingly to a flexible representation of information that can easily get updated by seamlessly including new data, entity types, and semantic relations [117, 26]. For this reason, many high-profile conversational agents take now advantage of large-scale knowledge graphs, such as the Google Knowledge Graph and the Alexa Knowledge Graph.

Recently, the focus shifted to the creation of sophisticated conversational agents that took advantage of transformers. GPT-2 (Generative Pre-trained Transformer) [125], GPT-3 [40], and the recent GPT-4 [114] are three examples in this field. GPT-3 was released in 2020 and became one of the largest language models to date, with 175 billion parameters. It was trained on a large corpus of 45 terabytes of text data from the internet, including books, articles, and websites, and can perform a wide range of natural language tasks, such as language translation, summarisation, and question-answering. GPT-4, is the next iteration of the GPT that is used in ChatGPT, but the exact details of its architecture and training data have not been disclosed.

In the last few months, GPT models have been used to power several prototypical chatbots targeted at the scholarly domain, such as Scite[4], Elicit[5], and CoreGPT[6].

[4] Scite - https://scite.ai/

[5] Elicit - https://elicit.org/

[6] CoreGPT - https://tinyurl.com/mvrk2z4x

These systems aim to assist users with a variety of tasks, such as identifying trends in the literature, choosing a venue for sharing their work, finding suitable collaborators, searching relevant articles, and more. However, it is not clear yet to what extent these new solutions can produce accurate answers about the academic landscape.

Although, several other domains have been affected by the introduction of chatbots, to the best of our knowledge, we still lack chatbots able to target the scholarly knowledge domain, and support the several stakeholders in this space, such as researchers, students, research policymakers, and companies. For instance, these solutions could support users in analysing trends in the literature, choosing a venue in which to disseminate their work, finding possible collaborators, identifying relevant articles, and so on. The architecture and the prototype presented in this paper aim at addressing this gap.

4.3 AIDA-Bot

The architecture of AIDA-Bot comprises two primary modules: Question Understanding and Response Generator.

The Question Understanding module is responsible for analyzing user inputs with the objective of identifying one of four predefined query types: count, list, describe, and compare. Its task is to transform the user's question into a formal query that can be applied to the knowledge graph. Notably, AIDA-Bot has enhanced capabilities, allowing for complex queries that can incorporate up to three filters. For instance, it can handle queries like, "List the top five papers about computer vision and machine learning written by researchers from the University of Cambridge." This represents an improvement over the previous version, which only supported a single condition.

In this module, a set of key terms is extracted from the user input and searched within the AIDA Knowledge Graph (AIDA KG) to identify relevant entities and their respective types. These extracted entities are then utilized to generate a set of pertinent questions that the system can automatically translate into queries suitable for the knowledge graph. The system also calculates the similarity between the user's input question and the set of generated questions. This approach enables the detection of a wide range of question formulations, encompassing various linguistic expressions.

If the similarity score between the user input and the most similar generated question exceeds a threshold, the Response Generator module uses a template to translate the latter to a query on AIDA KG and retrieves the relevant information. Otherwise, the system retrieves from AIDA KG the set of articles containing in the title or the abstract the key terms extracted from the user question. It then applies a question-answering model to produce a response based on the articles.

In the following, we describe the two modules in detail and provide more information on the adopted transformer models.

4.3.1 Question Understanding

The Question Understanding module is responsible for parsing the input query and utilizes named-entity recognition (NER) to identify key terms. These key terms encompass nouns, noun phrases, named entities, and compound expressions enclosed in quotes. To achieve this information extraction task, the module employs spaCy[7], an open-source Python library designed for Natural Language Processing (NLP)[8].

In the system, users have the flexibility to employ compound expressions enclosed in quotation marks to specify an exact match, akin to how search engines function. To avoid redundancy and streamline the key terms, we take the following steps: 1) We remove nouns and noun phrases that appear within a named entity or an expression enclosed in quotation marks. This ensures that terms already contained within such entities are not duplicated in the key terms. 2) We discard words that suggest questions (e.g., "who", "what"), and terms that indicate an entity type (e.g., "papers", "articles", "citations"). This helps refine the key terms by excluding common words that don't contribute to query specificity.

For instance, consider the request: " *Count papers about mathematics and matrix algebra written by authors from 'French Institute for research in computer science and automation'* ". In this example, the key terms initially include words like *papers, mathematics, matrix, algebra, authors, French, Institute, research, computer, science, automation.* These terms are also present in noun phrases, such as *'mathematics', 'matrix algebra', 'authors', 'French Institute', 'research', 'computer science', 'automation'* resulting in their removal to avoid redundancy.

Moreover, words like *papers* and *authors* are removed as they represent entity types within the AIDA Knowledge Graph (AIDA KG). Additionally, any terms appearing within the quoted expression are also discarded. Consequently, the resulting key terms for this example would be: *'mathematics', 'matrix algebra'* and *"French Institute for research in computer science and automation".* These refined key terms are then used for subsequent query construction and information retrieval.

The key terms, once identified, are searched within the AIDA Knowledge Graph (AIDA KG) to retrieve the relevant entities and their corresponding types. In the previous example, all key terms would be found in AIDA KG, with "French Institute for research in computer science and automation" recognized as *organization* while 'mathematics' and 'matrix algebra' as *topic.*

Subsequently, the Question Understanding module leverages these obtained entities to generate a grammar for creating all compatible requests that can be translated into queries for the knowledge graph. In this context, a grammar refers to a set of production rules that outline how to construct valid sentences or queries. These rules define the permissible combinations of symbols or tokens and the sequence in which they should appear.

In our system, the grammar is dynamically generated using templates that in-

[7] Spacy - `https://spacy.io/`.

[8] Specifically, we adopted the "en_core_web_sm" model.

corporate placeholders filled with the identified entities and their associated types. Below, we provide an example of a simple template for each query type:

1. count < sub_c > {}
2. list the <super> {num} <sub_l> {}
3. describe {}
4. compare {} vs {}

where:

- <sub_c> = papers | authors | conferences | organizations | citations | journals
- <super> = top | most important | main | most cited
- <sub_l> = papers | authors | conferences | organizations | topics | journals

Curly parentheses can only be filled with instances from the AIDA KG. Variables in angular parenthesis (e.g., <sub_c>) can only be filled with the previously defined items (e.g., papers, authors, conferences, and so on). Additionally, synonyms for these items, as pre-defined in a list, may also be employed. For example, <sub_c> would match both the words “papers” and “articles”.

During the generation of the grammar, the system will produce all questions compatible with the set of detected entities. When considering the four templates defined above, if the system detects entities of type [“topics”, “conferences", “organizations”, “authors”, “journals”], it will produce a range of questions of types 1 and 2. The module produces types 1 and 2 queries with up to three identified instances, allowing users to specify queries with three filters. Whenever at least one element from [“authors”, “conferences”, “organizations”] is found, the system will produce queries of type 3. Whenever it detects two items of the same class, it will generate queries of type 4.

In practice, each question type is supported by multiple templates since the same type of question can appear in several forms. For example, *how many <sub_c> {}* is another template for the query type *count* and would support questions such as “How many papers are there about the semantic web and machine learning?”. Therefore, from a modest number of initial templates covering the four query types (15 in the current implementation) and a set of identified entities, AIDA-Bot can generate a large number of candidate questions. Current templates were derived from use cases specific to SN and further improved through iterative refinement based on user feedback. Since developing new templates requires limited effort, the system can be easily adapted to other domains.

Next, the system computes the similarity between the original user request and the questions generated by the grammar. This step enables us to recognise a wide variety of formulations pertaining to the same question. In practice, we encode both the user's input and the generated questions as sentence embeddings and then compute their cosine similarity. If the similarity score between the user input and the most similar generated question exceeds an empirically established threshold, the module designates the latter as the representative of the user query. As this question was derived from a template, the system knows how to translate it into a query on the knowledge graph.

Finally, the Question Understanding module sends all pertinent information for the next phase to the Response Generator, including key terms, entities, entity types, and query types.

4.3.2 Response Generator

The Response Generator within the system distinguishes between two primary scenarios.

- Matching Generated Query: If the user's request aligns with one of the previously generated queries, the Response Generator proceeds to produce the corresponding query. It then executes this query over the AIDA Knowledge Graph (AIDA KG) and retrieves the relevant data.

  To provide a natural language response, the Response Generator utilizes response templates tailored to each specific query type. These templates are designed to be populated with the relevant data retrieved from AIDA KG. Additionally, the templates undergo further refinement, including adjustments to singular and plural terms, ensuring grammatical correctness and overall coherence in the generated answer. This process results in coherent and contextually appropriate responses to user queries.

- User Request Not Matching Generated Queries: When the user's question fails to match one of the generated queries, the Response Generator module handles the user's request as an 'open question'. In such cases, the module strives to generate a response by employing a question-answering model that operates on both the user request and the abstracts of relevant articles. To facilitate this process, the module first retrieves from the AIDA Knowledge Graph (AIDA KG) the set of papers that contain key terms relevant to the user's query. If the query returns no papers, typically because the user request falls outside the system's scope or doesn't align with any existing articles, AIDA-Bot proactively requests the user to reformulate or modify their request for better clarity and relevance. However, when relevant papers are found, the module proceeds to select those whose abstracts exhibit the highest similarity to the user query. This selection process is facilitated by a transformer model designed specifically for assessing sentence similarity. Following the selection of relevant papers, a summarization model is applied to condense the lengthy abstracts into more concise text, making it easier for the question-answering model to process. Subsequently, the module utilizes a question-answering model to generate a response to the user's question based on the condensed information from the relevant papers. This response is further enriched by providing a brief bibliography that lists the relevant articles. Whenever possible, the bibliography includes the Digital Object Identifiers (DOIs) and links to open-access versions of the articles, enhancing the user's ability to access and explore the referenced literature. This approach ensures that the user receives a comprehensive response even when their query doesn't match any pre-defined queries.

4.3.3 Transformer Models

AIDA-Bot leverages transformer models for three primary tasks: 1) Assessing Text Similarity, which is used for measuring sentence similarity, AIDA-Bot utilizes the 'all-MiniLM-L6-v2' model, from the Sentence-Transformers library[9]. This choice was made due to the model's efficiency and compact size. Widely recognized as state-of-the-art technology, it is highly regarded for its effectiveness in addressing tasks related to Semantic Textual Similarity (Reimers et al., 2019). The Sentence-Transformers framework is employed to access and utilize this model conveniently. This framework provides a convenient package for accessing BERT-based models and their variants, such as RoBERTa, MPNet, and ALBERT. 2) Question Answering, the transformer model used is 'distilbert-base-cased-distilled-squad'[10] from Huggingface. While maintaining performance comparable to BERT, this model consumes less computing power. It achieves a 60% faster runtime while retaining 95% of BERT's performance. It was developed by distilling the BERT base with 40% fewer parameters than the standard *textitbert-base-uncased.* 3) Text Summarization, AIDA-Bot employs the 'sshleifer/distilbart-cnn-12-6'[11] model from Huggingface for text summarization tasks. This model is based on DistilBART models, which are created by removing the decoder layers from a Seq2Seq transformer and then fine-tuning to produce high-quality student models. The model consistently generates superior summaries for AIDA-Bot's use cases, and distilbart-cnn-12-6 is chosen due to its significantly lighter footprint.

These transformer models, carefully selected for their efficiency and performance, play a crucial role in enhancing the capabilities of AIDA-Bot across various natural language understanding and generation tasks.

[9] `https://www.sbert.net/docs/pretrained_models.html`

[10] `https://huggingface.co/distilbert-base-cased-distilled-squad`

[11] `https://huggingface.co/sshleifer/distilbart-cnn-12-6`

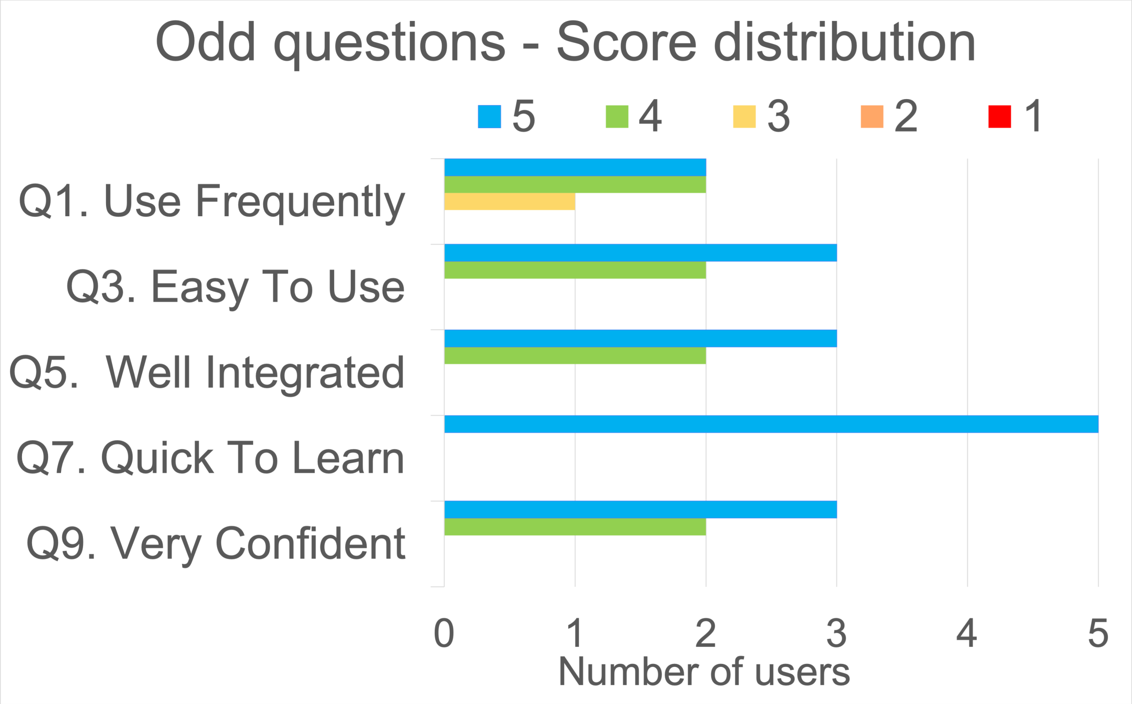


Figure 4.1: Odd questions. The higher the score, the better the system.

4.4 Evaluation

We conducted a user study involving five computer scientists from different institutions: the University of Cagliari (Italy), Gesis - Leibniz Institute for Social Science (Germany), and The Open University (UK). Their areas of expertise include Artificial Intelligence, Natural Language Processing, Semantic Web, Complex Networks, Data Science, and Big Data. This study aimed to assess the system's usability and gather feedback for improvements.

We began each session with a 15-minute presentation of AIDA-Bot and its capabilities. Then, we instructed the users to engage in an interactive session of about 45 minutes.

We asked them to complete a two-part survey describing their overall experience. The first section uses the standard *System Usability Scale* (SUS) questionnaire to assess the usability of AIDA-Bot. The second section includes five open questions regarding the strengths, weaknesses, and general feedback about AIDA-Bot. In what follows, we describe the outcome of these surveys.

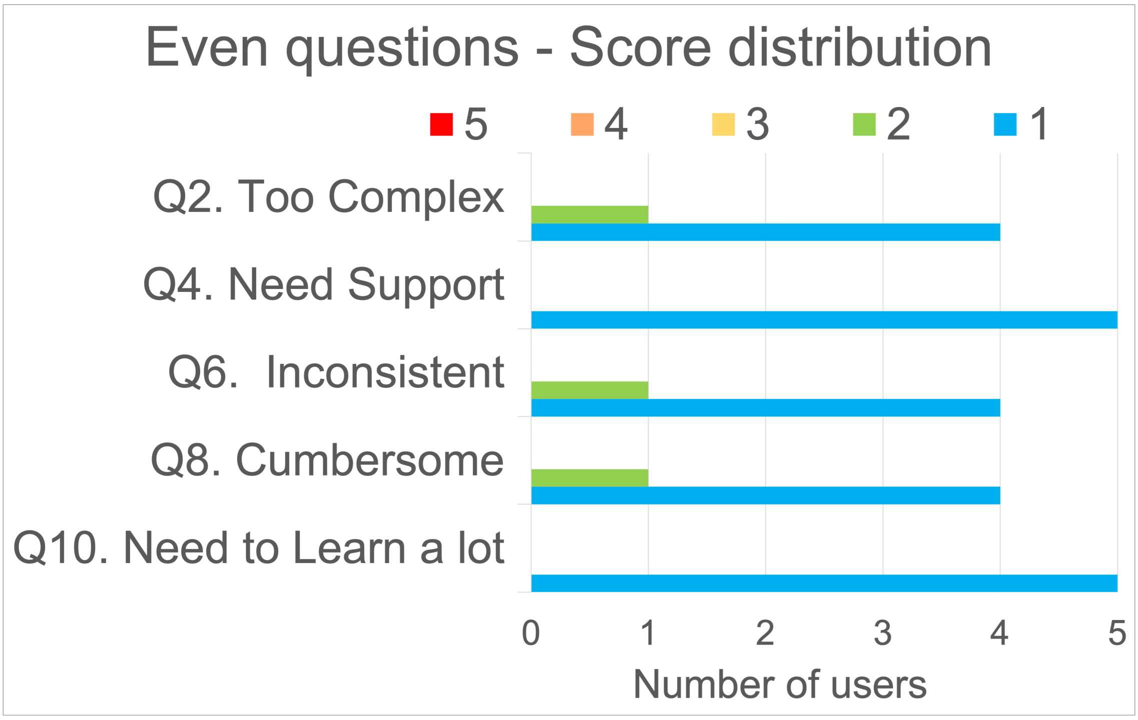


Figure 4.2: Even questions. The lower the score, the better the system.

SUS questionnaire.

The SUS questionnaire[12] provided excellent results, scoring 93.5/100, which is equivalent to an A grade, placing the AIDA-Bot in the 95 percentile rank[13].

Figures 4.1 and 4.2 depict the distribution of scores provided by users. Figure 4.1 displays the results for positive questions (odd-numbered), while Figure 4.2 presents feedback for negative questions (even-numbered).

Based on user feedback, AIDA-Bot received favorable ratings for usability. Users found it easy to use, with an average score of 4.6 ± 0.5. They also noted that its features were well-integrated, with a score of 4.6 ± 0.5. Users found it to be straightforward, with a low complexity rating of 1.2 ± 0.4, and indicated that they would not require assistance in using it in the future, with a score of 1.0 ± 0.0. Additionally, the System Usability Scale (SUS) results indicated that users felt highly confident while using the system (4.6 ± 0.5) and expressed a willingness to use it frequently (4.2 ± 0.8).

Open Questions.

In this section, we summarise the answers to the open questions.

[12] SUS Questionnaire Questions: https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html

[13] Interpreting a SUS score - https://measuringu.com/interpret-sus-score/

Q1. What are the main strengths of AIDA-Bot? User feedback highlighted several strengths of AIDA-Bot. Three users appreciated the system's simplicity and its ability to swiftly provide all the necessary information. One user found the system's capability to retrieve and explore scholarly information particularly valuable. It eliminates the need to search internal databases or the web, streamlining the research process. Another user identified one of the primary strengths as AIDA-Bot's ability to compare entities using predefined metrics, facilitating data-driven comparisons and analysis.

Q2. What are the main weaknesses of AIDA-Bot? The primary weakness identified by most evaluators pertains to the system's response time, with the current prototype occasionally taking several seconds to provide an answer. Additionally, two users expressed reservations about the quality of open-ended responses, citing occasional instances where the formulation appeared peculiar. It's worth noting that these concerns may arise from heightened user expectations influenced by the recent release of advanced GPT models, which have set a high standard for text generation quality.

Q3. Can you think of any additional features to be included in AIDA-Bot? The users offered insightful suggestions for enhancing the system. These include, i) a feature that allows the system to generate a bibliography based on user inputs, ii) an improvement in the system's approach to answering open-ended questions by incorporating GPT-like models, iii) The ability for the system to remember and reference what the user said earlier in the conversation.

Q4. Can you think of any additional types of queries for AIDA-Bot? The users provided several suggestions to enhance the system's functionality. First, they proposed enabling the system to answer arbitrary questions about the content of a specific paper. Second, they recommended expanding the system's entity comparison capabilities by allowing users to define arbitrary metrics for comparison. Lastly, they suggested incorporating a feature that would facilitate the identification of articles referencing specific analysis techniques, algorithms, or datasets. These enhancements would broaden the system's utility and versatility for users in various domains.

Q5. What would you add to increase the accuracy/comprehensiveness of the information returned by AIDA-Bot? Two users suggested improving the entity detection methodology by considering the complete conversational context, which would include previous messages and a user model. Additionally, one user recommended utilizing the full text of research papers, rather than just abstracts, to enhance the precision and comprehensiveness of extracted information.

In summary, the user study demonstrated that AIDA-Bot is highly usable and perceived as a valuable tool for providing accurate information about the research landscape. However, the emergence of modern GPT models has raised user expectations regarding the utilization of contextual information to comprehensively comprehend queries and generate highly coherent open-domain responses in real time. Although AIDA-Bot is designed specifically for answering questions about the research landscape, it may be beneficial to integrate some of these new solutions. The primary challenge going forward is to do so without compromising the accuracy of the resulting analytics or deviating from the verifiable information in the knowledge graph.

Chapter 5

Link Prediction

5.1 Introduction

The technology of Knowledge Graphs (KGs) empowered by graph-based knowledge representation brought an evolutionary change in a range of AI tasks. As a consequence, many application domains in science, industry, and different enterprises use KGs for data management. However, a challenge with KGs is that, despite the presence of millions of triples, capturing complete knowledge from the real world is almost impossible, even for specific application domains. Therefore, KGs usually remain incomplete.

Scientific research is one of the major domains for the application of KGs. In the last years, we saw the emergence of several KGs describing research outputs, such as Microsoft Academic Graph[1] [167], Scopus[2], Semantic Scholar[3], Aminer [179], Core [79], OpenCitations [123], Dimensions[4], Open Research Knowledge Graph[5] [71], and others. These solutions are crucial for performing large-scale bibliometric studies, informing funding agencies and research policymakers, supporting a variety of intelligent systems for querying the scientific literature, identifying research topics, suggesting relevant articles and experts, detecting research trends, and so on. Their usefulness and, consequently, our ability to assess research dynamics, are however crucially limited by their incompleteness. Even basic metadata such as affiliations, organization types, references, research topics, and conferences are often missing, noisy, or not properly disambiguated. Therefore, apparently simple tasks such as identifying the affiliation and the country of origin of a publication still require a large amount of manual data cleaning [93].

Traditionally, data integration methods have been applied to solve data incompleteness in the context of databases and repositories. However, when completing and refining large KGs, it is crucial to adopt scalable and automatic approaches. Among the many possible graph completion methods, Knowledge Graph Embedding (KGE) models have recently gained a lot of attention. KGEs learn representations of graph nodes and edges with the goal of predicting links between existing entities. Embedding models have been in practical use for various types of KGs in different domains, including digital libraries [175], biomedical [86], and social media [152].

The motivating scenario for this work was supplied by the Academia/Industry DynAmics (AIDA)[6] Knowledge Graph [21], a resource that was designed for studying the relationship between academia and industry and for supporting systems for predicting research dynamics. The version of AIDA used in this work integrates the metadata about 21M publications from Microsoft Academic Graph (MAG) and 8M patents from Dimensions in the field of Computer Science. In this resource, docu-

[1] Microsoft Academic Graph - `http://aka.ms/microsoft-academic`

[2] Scopus - `https://www.scopus.com/`

[3] Semantic Scholar - `https://www.semanticscholar.org/`

[4] Dimensions - `https://www.dimensions.ai/`

[5] ORKG - `https://www.orkg.org/orkg/`

[6] AIDA - `http://aida.kmi.open.ac.uk`

ments are categorized according to their research topics drawn from the Computer Science Ontology (CSO)[7] [140] and classified with their authors' affiliation types on the Global Research Identifier Database (GRID)[8] (e.g., 'Education', 'Company', 'Government', 'Nonprofit'). This solution enables analysing the evolution of research topics across academia, industry, government institutions, and other organizations. For instance, it allows us to detect that a specific topic, originally introduced by academia, has been recently adopted by industry. It can also support systems for predicting the impact of specific research efforts on the industrial sector [136] and the evolution of technologies [115]. Nevertheless, only 5.1M out of the 21M articles could be mapped to a GRID and characterized according to their affiliation type. Therefore, more than 75% of the publications are missing this critical information, significantly reducing the scope and accuracy of the resulting analytics. In order to show that our approach can be applied to fields with very different characteristics, we also use it to complete the Fields of Study, which is a collection of terms from multiple disciplines utilized to index the articles in MAG. Indeed, the completeness of the set of terms associated with a paper varies a lot and depends on the quality and style of the abstract, which in turn is often parsed from online PDFs, leading to mistakes and missing content. This in turn hinders our ability to understand the research concepts associated with the paper and to obtain comprehensive analytics.

Completing the affiliation types and the Fields of Study is crucial for improving the overall quality of these knowledge graphs and a very good practical use case for link prediction.

We evaluated Trans4E against several alternative models (TransE, RotatE, QuatE, ComplEx) on the task of link prediction on AIDA, MAG, and four other well-known benchmarks (FB15K, FB15k-237, WN18, and WN18RR).

The experiments showed that Trans4E outperforms the other approaches in the case of N to M relations with N $\gg$ M and yields very competitive results in all the other cases, in particular when using low embedding dimensions. The ability to solve the N $\gg$ M issue and to perform well even when adopting small embedding dimensions makes Trans4E particularly apt for handling large scale knowledge graphs that describe millions of entities of the same type (e.g., documents, persons).

In summary, the contributions of our work are the following:

- We propose Trans4E, a new embedding model specifically designed to provide link prediction for large-scale KGs presenting N to M relations with N $\gg$ M.
- We apply Trans4E on a real word scenario that involves completing affiliation types and Fields of Study (N $\gg$ M relations) in AIDA and MAG.
- We further evaluate our approach on four well-known benchmarks (FB15k, FB15k-237, WN18, and WN18RR), showing that Trans4E yields competitive performances in several configurations.

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

[8] GRID - `https://www.grid.ac/`

The rest of the chapter is organised as follows. In Section5.2 , we review the literature on current embedding models for data completion. In Section5.3, we present the open issues. In Section5.4, we describe Trans4E. Section5.5 reports the evaluation of the model versus alternative solutions.

5.2 Background

In this section, we will first review the graph embedding models and their application to link prediction.

5.2.1 Knowledge Graph Embeddings

In this paragraph, we introduce the definitions required to understand our approach.

Embedding Vectors. Let the knowledge graph be $KG = (\mathcal{E}, \mathcal{R}, \mathcal{T})$, where $\mathcal{E}$ is the set of entities (nodes) in the graph, $\mathcal{R}$ is the set of all relations (edges), and $\mathcal{T}$ is the set of all triples in the graph in the form of (h, r, t), e.g., *(Berlin, CapitalOf, Germany)*. KGE models are applied to KGs for link prediction by measuring the degree of correctness of a triple. To do so, a KGE model aims at mapping each entity and relation of the graph into a vector space (shown as $(\mathbf{h}, \mathbf{r}, \mathbf{t})$, $\mathbf{h}, \mathbf{r}, \mathbf{t} \in \mathbb{R}^d$), where d is the embedding dimension of each vector. By h_i, we refer to the i-th element of the vector $\mathbf{h}$ where i ranges in $\{1, \dots, d\}$. The vector representation of the entities and the relations in a KG are the actual embeddings.

Score Function. Using this representation, the plausibility of the triples is then assessed by the scoring function $f(\mathbf{h}, \mathbf{r}, \mathbf{t})$ of the applied KGE model. If a triple is more plausible, its score should be higher. For example, $f($*Berlin, CapitalOf, Germany*$)$ should be higher than $f($*Berlin, CapitalOf, France*$)$.

Negative Sampling. Traditional machine learning approaches typically involve training on both positive and negative samples. However, in the context of Knowledge Graphs (KGs), all the triples within them are inherently considered as true statements. This unique characteristic necessitates the introduction of negative samples into the training process of Knowledge Graph Embeddings (KGEs). In this chapter, we employ the technique known as Adversarial Negative Sampling (adv) for this purpose. This approach entails generating a set of negative samples derived from a given triple (h, r, t) by employing a probabilistic algorithm that replaces either the subject h or the object t with a random entity (h' or t') drawn from the set of entities $\mathcal{E}$.

Loss Function. Due to the initial phase of the learning process, embedding vectors are initialized with random values, the scores of the triples for positive and negative samples are also random. To address this, an optimization process is employed to adapt the embeddings such that positive samples yield higher scores compared to negative ones. This optimization is achieved by minimizing a loss function denoted as $\mathcal{L}$. The commonly employed method for optimizing the loss function is Stochastic Gradient Descent (SGD).

N to M Relations. As mentioned above, given a relation r, the representation of facts in triple form is (h, r, t). Depending on the type of a relation and its meaning, for a fixed head (say h_1), there are at most M possible tails connected to the head, i.e. $\{(h_1, r, t_1), (h_1, r, t_2)\}, \ldots, (h_1, r, t_M)$. Similarly, for a fixed tail, (say t_1), there are at most N possible head entity, i.e. $\{(h_1, r, t_1), (h_2, r, t_1), \ldots, (h_N, r, t_1)\}$. There are four cases that may arise for a relations which connects a different number of heads and tails: a) both N and M are small, b) both M and N are large, c) N is small and M is large, and d) N is large and M is small. The latter is the focus of this chapter. For example, in the AIDA knowledge graph the "hasType" relations connects a very large number of head entities (5.1M articles) to only 8 tail entities (the GRID types).

5.2.2 Review of State-of-the-art KGEs

Here we summarize some of the most used existing models focusing in particular on their scoring function.

TransE

[36] is one of the early embedding models and is well known for its outstanding performance and simplicity. It is a solid baseline that can still outperform many of the most recent and complex KGEs [65]. The idea of the TransE model is to enforce embedding of entities and relations in a positive triple (h, r, t) to satisfy the following equality:

$$\mathbf{h} + \mathbf{r} \approx \mathbf{t} \tag{5.1}$$

where $\mathbf{h}$, $\mathbf{r}$ and $\mathbf{t}$ are the embedding vectors of head, relation, and tail, respectively. TransE model defines the following scoring function:

$$f_r(h, t) = -\|\mathbf{h} + \mathbf{r} - \mathbf{t}\| \tag{5.2}$$

RotatE

[155] is a model designed to transform the head entity to the tail entity by using the relation rotation. This model embeds entities and relations in complex space. If we constrain the norm of entity vectors, this model would be reduced to TransE. The scoring function of RotatE is

$$f_r(h, t) = -\|\mathbf{h} \circ \mathbf{r} - \mathbf{t}\| \tag{5.3}$$

in which $\circ$ is the element-wise product. Rotate is one of the recent state-of-the-art models which is leading the accuracy competition among KGEs [155].

ComplEx

[160] is a semantic matching model, which assesses the plausibility of facts by considering the similarity of their latent representations. In other words, it is assumed that similar entities have common characteristics, i.e. are connected through similar relationships [109, 168]. In ComplEx the entities are embedded in the complex space. The score function of ComplEx is given as follows:

$$f(h,t) = \Re(\mathbf{h}^T \operatorname{diag}(\mathbf{r})\, \bar{\mathbf{t}})$$

in which $\bar{\mathbf{t}}$ is the conjugate of the vector $\mathbf{t}$ and $\Re$ returns the real part of the complex number.

QuatE

[177] model relations in the quaternion space. Similarly to RotatE, QuatE represents a relation as a rotation. However, a rotation in quaternion space is more expressive than a rotation in complex space. A product of two quaternions $Q_1 \otimes Q_2$ is equivalent to first scaling Q_1 by magnitude $|Q_2|$ and then rotating it in four dimensions. QuatE finds a mapping $\mathcal{E} \to \mathbb{H}^d$, where an entity h is represented by a quaternion vector $\mathbf{h} = a_h + b_h\mathbf{i} + c_h\mathbf{j} + d_h\mathbf{k}$, with $a_h, b_h, c_h, d_h \in \mathbb{R}^d$.

The scoring function is computed as follows:

$$\phi(h,r,t) = \mathbf{h}' \cdot \mathbf{t} = \langle a'_h, a_t \rangle + \langle b'_h, b_t \rangle + \langle c'_h, c_t \rangle + \langle d'_h, d_t \rangle \tag{5.4}$$

where $\langle \cdot, \cdot \ \rangle$ is the inner product. $\mathbf{h}'$ is computed by first normalizing the relation embedding $\mathbf{r} = p_r + q_r\mathbf{i} + u_r\mathbf{j} + v_r\mathbf{k}$ to a unit quaternion:

$$\mathbf{r}^{(n)} = \frac{r}{|r|} = \frac{p_r + q_r\mathbf{i} + u_r\mathbf{j} + v_r\mathbf{k}}{\sqrt{p_r^2 + q_r^2 + u_r^2 + v_r^2}} \tag{5.5}$$

and then computing the Hamiltonian product between $\mathbf{r}^{(n)}$ and $\mathbf{h} = a_h + b_h\mathbf{i} + c_h\mathbf{j} + d_h\mathbf{k}$:

$$\begin{aligned}\mathbf{h}' = \mathbf{h} \otimes \mathbf{r}^{(n)} := {} & (a_h \circ p - b_h \circ q - c_h \circ u - d_h \circ v) \\ & + (a_h \circ q + b_h \circ p + c_h \circ v - d_h \circ u)\, \mathbf{i} \\ & + (a_h \circ u - b_h \circ v + c_h \circ p + d_h \circ q)\, \mathbf{j} \\ & + (a_h \circ v + b_h \circ u - c_h \circ q + d_h \circ p)\, \mathbf{k}\end{aligned} \tag{5.6}$$

5.3 Open Issues

Traditionally, data integration methods have been applied to solve data incompleteness in the context of databases and repositories. However, when completing and refining large KGs, it is crucial to adopt scalable and automatic approaches. Among the many possible graph completion methods, Knowledge Graph Embedding (KGE) models have recently gained a lot of attention. KGEs learn representations of graph nodes and edges with the goal of predicting links between existing entities. Embedding models have been in practical use for various types of KGs in different domains, including digital libraries [175], biomedical [86], and social media [152].

However, the specific characteristics of scholarly KGs poses important challenges for link prediction methods based on KGE models [36, 155, 160, 177, 169, 159, 105]. One crucial aspect is the presence of several N to M relations with N$\gg$M. Given a triple (h, r, t), this situation arises when the cardinality of the entities in the head position (h) for a certain relation (r) is much higher than the one of the entities in the tail position (t). This is the case for most scholarly knowledge graphs [123, 167, 9, 179, 78] that usually categorize millions of documents (e.g., papers, patents) according to a relatively small set of categories (e.g., topics, affiliation kinds, countries, chemical compounds).

Current KGE models lack the ability to handle effectively these kinds of relations since they are unable to assign to each entity a well distinct embedding vector in a low dimensional space. As a result, link prediction and node classification techniques that exploit these embeddings tend to perform poorly.

To address this problem, in this chapter we will analyze Trans4E, a new embedding model specifically designed to support link prediction for KGs which present N to M relations with N$\gg$M. Specifically, Trans4E tackles the issue by providing a larger number of possible vectors ($8^d - 1$, where d is the embedding dimension) to be assigned to entities involved in N to M relations. Trans4E enables the generation of a well distinct vector for each entity even when using small embedding dimensions.

5.4 Trans4E

5.4.1 The Trans4E Model

Trans4E is a novel KGE model designed to effectively handle KGs which include N to M relations with N$\gg$M.

In this section, we show that the capacity of this model for a given relation (e.g., *hasGRIDType*, *hasTopic*) and the corresponding tail entity (e.g., *type* or *topic*) is 8^d, which allows to generate a distinct vector for each entity (e.g., a specific paper) even when using small embedding dimensions.

Here we introduce the core formulation of the score function of Trans4E.

Trans4E maps the entities of the graph via relations in Quaternion vector space $\mathbb{H}^d$.

Concretely, given a triple of the form (h, r, t), our model follows the following steps:

(a) The head entity vector ($\mathbf{h} \in \mathbb{H}^d$) is rotated by $\mathbf{r}_\theta$ degrees in quaternion space i.e. $\mathbf{h}_{\theta_r} = \mathbf{h} \otimes \mathbf{r}_\theta$. $\otimes$ is an element-wise Hamilton product between two quaternion vectors.

(b) The rotated head i.e. $\mathbf{h}_{\theta_r}$ is translated by the relation embedding vector $\mathbf{r}$ to get $\mathbf{h}_r = \mathbf{h}_{\theta_r} + \mathbf{r}$.

(c) The translated head embedding vector should meet the tail embedding vector i.e. $\mathbf{h}_r \approx \eta_h \otimes \mathbf{t}$ for a positive sample (h, r, t). $\mathbf{t} \in \mathbb{H}^d$. However, there is a possibility that the transformed vector of the head is not exactly meeting the tail. In order to solve this problem, we could use $\eta_h = [\eta_{h1}, \dots, \eta_{hd}] \in \mathbb{H}^d$, which is a mapping regularizer.

Following the mentioned steps, we define the score function as:

$$f(\mathbf{h}, \mathbf{r}, \mathbf{t}) = -\|\mathbf{h}_r - \eta_h \otimes \mathbf{t}\|. \tag{5.7}$$

The score function returns a low value if the triple is false i.e. $\mathbf{h}_r \neq \eta_h \otimes \mathbf{t}$ and returns high value (close to zero) if the triple is true i.e. $\mathbf{h}_r \approx \eta_h \otimes \mathbf{t}$. In this way, we measure the plausibility of each triple (h, r, t).

Furthermore, two regularized variants of the Trans4E model have been introduced for this study. The first one is referred to as Trans4EReg1, which is a regularization of the Trans4E model. Trans4EReg1 incorporates relation-specific head rotation and tail mapping regularization. The second variant is Trans4EReg2, another regularization of Trans4E. Trans4EReg2 features a relation-specific rotation applied to the tail side, in addition to the inclusion of relation-specific head rotation and tail mapping regularization.

5.4.2 Link Prediction on N to M Relations

Here we show that Trans4E provides a higher capacity with fewer limitations than other models.

Given a relation r (e.g., *hasGRIDType*) and a tail t (e.g., 'Education'), the following constraints are applied for each of the resulting triples:

$$\begin{cases} \mathbf{h}_{1\theta_{ri}} + \mathbf{r}_i = \eta_{h_{1i}} \otimes \mathbf{t}_i, \\ \mathbf{h}_{2\theta_{ri}} + \mathbf{r}_i = \eta_{h_{2i}} \otimes \mathbf{t}_i, \\ \vdots \\ \mathbf{h}_{N\theta_{ri}} + \mathbf{r}_i = \eta_{h_{Ni}} \otimes \mathbf{t}_i, \qquad i = 1, \dots, d. \end{cases} \tag{5.8}$$

We can rewrite the Hamilton product as 4-dimensional matrix-vector product:

$$
\begin{aligned}
&\mathbf{h}_{\theta_{ri}} = \mathbf{h}_i \otimes \mathbf{r}_{\theta_i} = \\
&\begin{bmatrix} a_{r_\theta} & -b_{r_\theta} & -c_{r_\theta} & -d_{r_\theta} \\ b_{r_\theta} & a_{r_\theta} & -d_{r_\theta} & c_{r_\theta} \\ c_{r_\theta} & d_{r_\theta} & a_{r_\theta} & -b_{r_\theta} \\ -d_{r_\theta} & -c_{r_\theta} & b_{r_\theta} & a_{r_\theta} \end{bmatrix} \begin{bmatrix} a_h \\ b_h \\ c_h \\ d_h \end{bmatrix} = \mathcal{H}_i \vec{h}_i.
\end{aligned}
\tag{5.9}
$$

Without loss of generality, we assume that the embedding of the relation translation $\mathbf{r}_i$ is zero and $\eta_{h_{pi}}$ is a real value. In this way, we can write the above system of equations in the following form:

$$
\begin{cases}
\mathcal{H}_i \vec{h}_{1i} = \eta_{h_{1i}} \vec{t}_i \\
\mathcal{H}_i \vec{h}_{2i} = \eta_{h_{2i}} \vec{t}_i \\
\vdots \\
\mathcal{H}_i \vec{h}_{1i} = \eta_{h_{Ni}} \vec{t}_i, \qquad i = 1, \dots, d.
\end{cases}
\tag{5.10}
$$

It's important to note that the matrix $\mathcal{H}_i$ is a 4 × 4 matrix with four distinct eigenvalues and their corresponding eigenvectors. As a result, we can express the relationship as $\mathcal{H}_i \vec{h}_{pi} = \lambda_{h_{pi}} \vec{h}_{pi} = \eta_{h_{pi}} \vec{t}_{pi}$. When $\lambda_{h_{pi}}$ equals $\eta_{h_{pi}}$, the ith dimension of the head and tail vectors will be identical; otherwise, they will differ. Consequently, in each dimension, we have a total of 8 possible options to assign to the head entity vector, as we have 4 distinct eigenvectors with two cases: one where the head and tail are equal and the other where they are not.

Considering that we utilize vectors of dimensionality d, there are a total of $8^d - 1$ possible distinct vectors to be assigned to entities that appear in the head, such as articles in AIDA. Consequently, the model's capacity becomes $8^d - 1$, offering a more expansive space compared to the TransE and RotatE models.

In Section 5.5, we will show the advantages of this solution by comparing it against alternative models.

5.5 Evaluation

We compared Trans4E against four alternative embedding models: TransE, RotatE, ComplEX, and QuatE.

5.5.1 Evaluation Datasets

We ran the experiments on a portion of the knowledge graph *AIDA+MAG* including 68,906 entities and 180K triples. Specifically, we considered the following entities: publication IDs, authors, affiliation organizations, topics, publication types, conference editions, conference series, journals, years, countries, and references.

In this subset, the *hasGRIDType* relation includes about 5k entities (research papers) in the head position and 7 entities as tail ('Education', 'Company', 'Government', 'Healthcare', 'Nonprofit', 'Facility', and 'Other').

Regarding the *hasTopic* relation, the highest number of research articles associated to a topic is 4,659, while the highest number of topics associated to research articles is only 13.

We split the datasets into train (80%), test (10%), and validation (10%) sets. Additionally, we evaluated the performance of our model on four benchmarks: FB15K (14,951 entities and 1,345 relations), FB15k-237 (14,451 entities and 237 relations), WN18 (40,943 entities and 18 relations), and WN18RR (40,943 entities and 13 relations).

5.5.2 Evaluation Criteria

In this section we discuss the criteria that we considered for the evaluation.

Performance Metrics. The standard evaluating metrics for the performance of KGEs are: Mean Rank (MR), Mean Reciprocal Rank (MRR) and Hits@k (k=1, 3, 10) [168].

MR is the average rank of correct triples in the test set. In order to compute it, we generate two sets of triples, $S_h = (h, r, ?)$ and $S_t = (?, r, t)$, by corrupting each test triple (h, r, t). After this step, the scores of all the triples in S_h, S_t are computed and the triples are sorted. The rank (r_h, r_t) of the original triple (i.e. (h, r, t)) is then computed in both sets S_h, and S_t. For any triple, r_h is the notation for the right ranks and r_t for the left ranks. The rank of the example triple of (h, r, t) is computed as $rank = \frac{r_h + r_t}{2}$. If we assume $rank_i$ to be the rank of the $i-$th triple in the test set obtained by a KGE model, then the MR and the MRR are obtained as follows:

$$MR = \sum_i rank_i,$$

$$MRR = \sum_i \frac{1}{rank_i}.$$

For the evaluation on *hasGRIDType* and *hasTopic* relations, we only corrupted the tail of the relations and replaced it with all the entities in the KG.

The $Hits@K$, for k = 1, 3, 10 ..., is one of the standard link prediction measurements. By considering the percentage of the triples for which $rank_i$ is equal or smaller than k, we computed the $Hits@K$. MR, the average MRR, Hits@1, Hits@3, and Hits@10 are reported in Tables 2-6.

Dimension and KG Scale. Although the performance measures of a machine learning model are important criteria for evaluation, the dimension of the embedding vectors is specifically important for KGE models, which are supposed to be used in the real-world large-scale KGs. Indeed, an embedding with very large dimensions may be unfeasible in most practical settings.

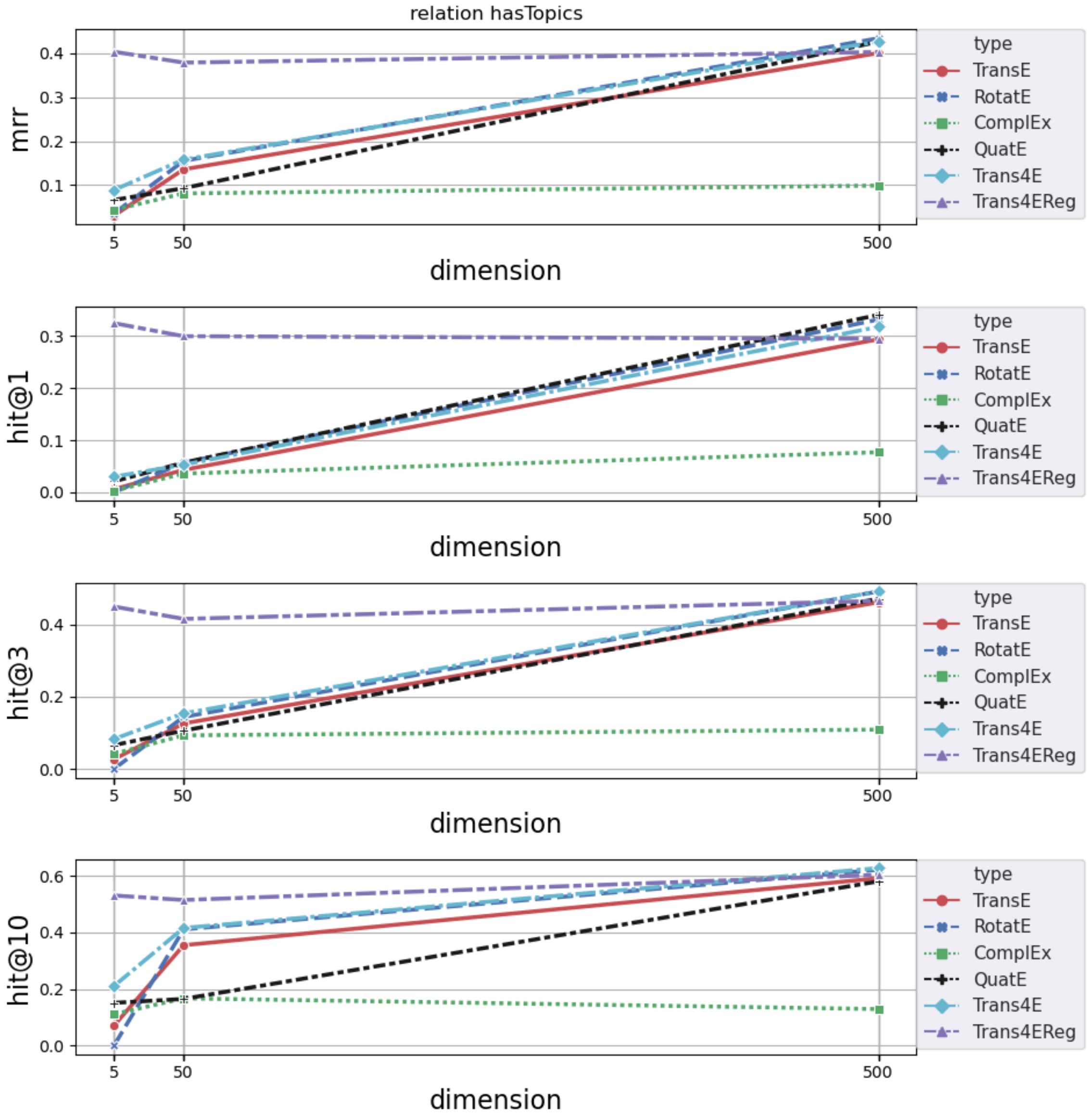


Figure 5.1: hasTopic for dimension 5,50 and 500

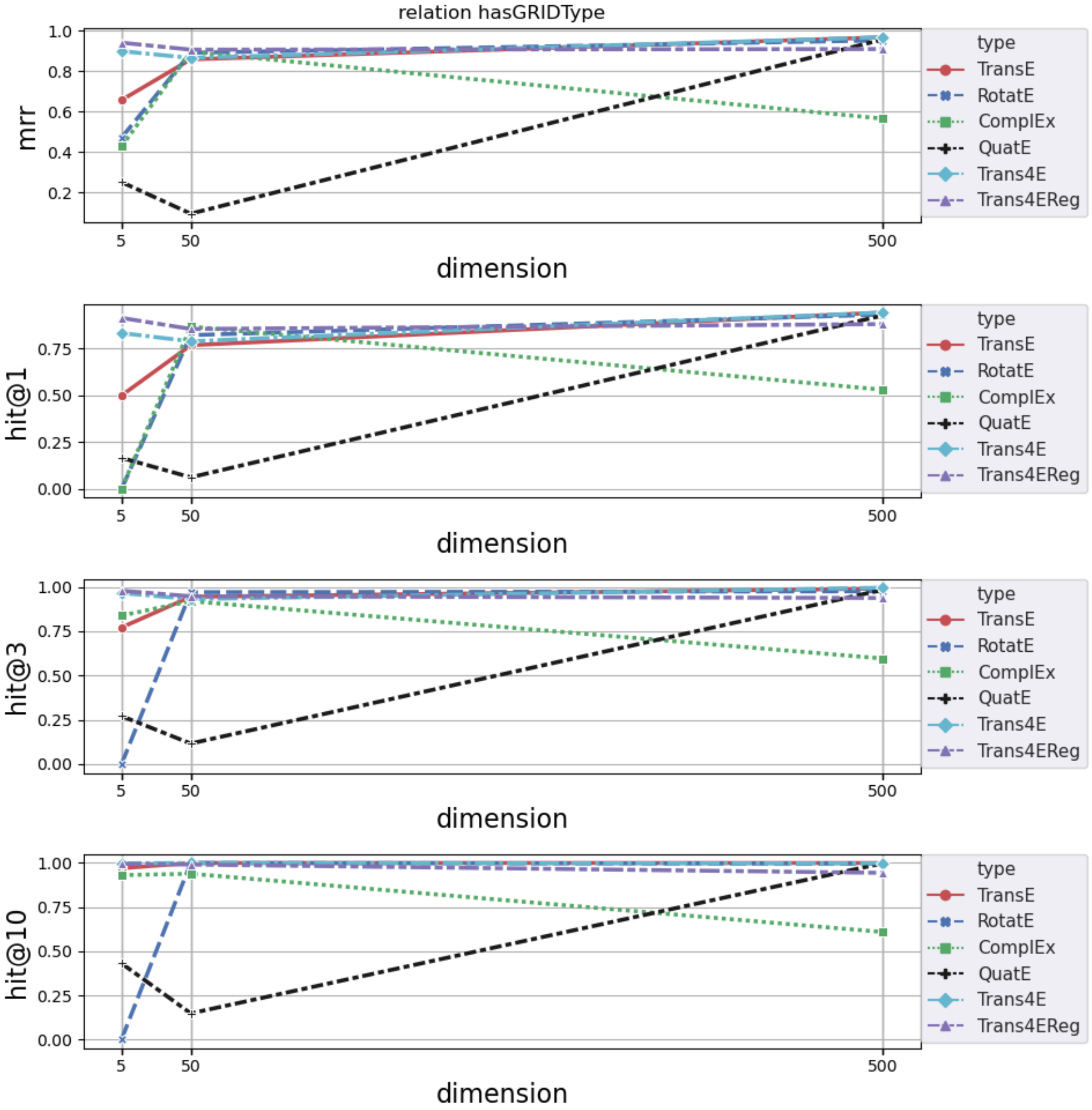


Figure 5.2: hasGRIDType for dimension 5,50 and 500

Table 5.1: Performance of KGEs on AIDA for Dimension 5

Model Type	hasTopic					hasGRIDType				
	MR	MRR	Hits@1	Hits@3	Hits@10	MR	MRR	Hits@1	Hits@3	Hits@10
TransE	3785	0.031	0.006	0.027	0.071	6	0.658	0.500	0.771	0.970
RotatE	4749	0.036	0.000	0.001	0.008	38	0.472	0.000	0.000	0.001
QuatE	4862	0.066	0.021	0.066	0.151	159	0.252	0.166	0.271	0.431
ComplEx	3726	0.044	0.003	0.042	0.111	6	0.429	0.001	0.838	0.931
Trans4EReg1	3007	0.403	0.325	0.450	0.531	1	0.941	0.915	0.978	0.995
Trans4EReg2	2047	0.401	0.325	0.445	0.528	1	0.956	0.928	0.985	0.988
Trans4E	2908	0.089	0.030	0.083	0.211	1	0.900	0.834	0.965	0.998

Table 5.2: Performance of KGEs on AIDA for Dimension 50

Model Type	hasTopic					hasGRIDType				
	MR	MRR	Hits@1	Hits@3	Hits@10	MR	MRR	Hits@1	Hits@3	Hits@10
TransE	3903	0.135	0.043	0.126	0.355	1	0.859	0.769	0.944	1.000
RotatE	3890	0.155	0.057	0.144	0.411	1	0.891	0.823	0.970	1.000
QuatE	1693	0.093	0.057	0.106	0.165	1718	0.096	0.062	0.116	0.148
ComplEx	7279	0.081	0.036	0.093	0.167	700	0.896	0.869	0.919	0.939
Trans4EReg1	2424	0.379	0.300	0.416	0.515	117	0.907	0.856	0.947	0.991
Trans4EReg2	3250	0.394	0.327	0.429	0.507	1	0.959	0.928	0.990	1.000
Trans4E	3842	0.158	0.053	0.154	0.416	1	0.866	0.790	0.931	1.000

Table 5.3: Performance of KGEs on AIDA for dimension 500.

Model Type	hasTopic					hasGRIDType				
	MR	MRR	Hits@1	Hits@3	Hits@10	MR	MRR	Hits@1	Hits@3	Hits@10
TransE	3982	0.400	0.294	0.462	0.592	1	0.968	0.944	0.990	1.000
RotatE	4407	0.433	0.332	0.492	0.622	1	0.953	0.933	0.975	0.996
QuatE	1353	0.426	0.341	0.472	0.581	1	0.957	0.928	0.983	0.998
ComplEx	5855	0.099	0.077	0.109	0.129	1566	0.566	0.531	0.596	0.609
Trans4EReg1	2040	0.402	0.295	0.466	0.604	233	0.910	0.882	0.937	0.944
Trans4EReg2	1942	0.424	0.325	0.482	0.602	34	0.955	0.931	0.978	0.990
Trans4E	3904	0.426	0.318	0.492	0.628	1	0.968	0.944	0.995	0.998

Therefore, we compared the performances of our model against state-of-the art models in a very low dimensional embedding. This was done to simulate a real-world application of KGEs on large scale KGs. Indeed, models which obtain satisfactory performances on a portion of a graph using a small vector size should also perform well when adopting a higher dimension on a larger portion of the same graph [106, 61].

5.5.3 Hyperparameter Setting

The development environment of our model is PyTorch[9]. In the experiments, we reshuffled the training set in each epoch, and generated 16 mini batches on the reshuffled samples. To determine the performances of our model in high and low dimensions, the embedding dimension (d) was set to $\{5, 50, 500\}$ in the experiments. The batch size (b) is considered as $\{256, 512\}$, the fixed margin γ is

[9] PyTorch - `https://pytorch.org/`

$\{2,3,4,5,10,15,20,30\}$ and learning rate as $\{0.001,0.01,0.05,0.1\}$ with a negative sample of 10. L_2 regularization coefficient is $\{0.000005,0.0000005\}$ for the models QuatE, Trans4EReg1, and Trans4EReg2. The best hyperparameter combination for Trans4E and Trans4EReg2 is $b=256$, $lr=0.1,\gamma=20$ and for Trans4EReg1 is $b=256$, $lr=0.001,\gamma=20$, and $d=500$ for all the models. For the regularized versions λ =0.000005.

5.5.4 Results and Discussions

In this section, we present the outcomes of our experiments. More specifically, the results of the evaluation for graph completion on AIDA+MAG are detailed in Section 5.5.4. Section 5.5.4 provides a performance comparison of Trans4E and various alternative methods across a range of standard benchmarks, including FB15k, FB15k-237, WN18, and WN18RR. Additionally, Section 5.4.3 investigates into the analysis of the representation of research topics, showcasing a study on the distribution of their embedding vectors.

Knowledge Graph Completion in AIDA+MAG

In this section we evaluate the performance of Trans4E versus alternative methods in completing the two relations *hasGRIDType* and *hasTopic* in AIDA+MAG.

Specifically, we compared Trans4E with TransE, RotatE, QuatE and ComplEx. We also included Trans4EReg1 and Trans4EReg2, the two regularized versions previously defined in Section 4.1.

Table 5.1 reports the performances of the seven models for dimension 5. Trans4EReg1 clearly outperforms all the other models for the *hasTopic* relations. Trans4EReg2 obtains the second-best performance. For instance, when considering the *hasTopic* relation, Trans4EReg1 and Trans4EReg2 yield 32.5% in Hits@1 while all the other solutions obtain less than 3%. For the *hasGRIDType* relations Trans4EReg2 outperforms all the others with a 92.8% in Hits@1. Moreover, Trans4EReg1, yields 91.5% in Hits@1 and Trans4E 83.4%, while the best of the other models is TransE with 50.0%.

RotatE performed surprisingly poorly on both the *hasTopic* and *hasGRIDType* relations, yielding 0% in Hits@1. It should be noted that during the testing phase, for each test triple $(p,\text{hasGRIDType},t)$, we systematically replaced the tail entity t with all entities present in the graph. Subsequently, we ranked the actual entities against the corrupted triples in this process. Notably, RotatE, with a dimensionality of 5, did not rank any typed entities among the top 10 occurrences. This observation implies that non-typed entities are ranked higher than typed entities in the corruption process. This phenomenon can be attributed to the constrained solution space of the RotatE model, a topic that is also explored in [107].

The overall accuracy for *hasGRIDType* is typically higher than *hasTopic*. For instance, Trans4EReg1 yields a Hits@10 of 99.5% for *hasGRIDType* and 53.1% for

Table 5.4: Performance of KGEs on FB15K and WN18.

Model Type		FB15k				WN18				
	MR	MRR	Hits@1	Hits@3	Hits@10	MR	MRR	Hits@1	Hits@3	Hits@10
TransE	–	0.463	0.297	0.578	0.749	–	0.495	0.113	0.888	0.943
RotatE	40	0.797	0.746	0.830	0.884	309	0.949	0.944	0.952	0.959
QuatE	35	0.742	0.658	0.805	0.881	349	0.942	0.927	0.952	0.960
Trans4E	47	0.767	0.681	0.834	0.892	175	0.950	0.944	0.953	0.960

Table 5.5: Performance of KGEs on FB15K-237 and WN18RR.

Model Type		FB15k-237				WN18RR				
	MR	MRR	Hits@1	Hits@3	Hits@10	MR	MRR	Hits@1	Hits@3	Hits@10
TransE	357	0.294	–	–	0.465	3384	0.226	–	–	0.501
RotatE	177	0.338	0.241	0.375	0.533	3340	0.476	0.428	0.492	0.571
QuatE	170	0.282	0.178	0.315	0.501	2272	0.303	0.179	0.386	0.530
Trans4E	158	0.332	0.236	0.366	0.527	1755	0.469	0.416	0.487	0.577

hasTopic. This is mainly due to the fact that the number of entities to be considered for *hasTopic* is much higher than that for *hasGRIDType*.

Overall, Trans4EReg1 seems to be the most suitable model for addressing large-scale KGs, where increasing the dimension of the model is too costly in computational terms.

Table 5.2 reports the performances of the models using dimensions 50. Trans4EReg1 and Trans4EReg2 outperforms all the models with regards to the *hasTopic* by a considerable margin (up to 10% improvement on Hits@10). When considering *hasGRIDType*, Trans4EReg2 obtains the best performances in all metrics, folowed by Trans4EReg1 and RotatE. Due to the overfitting, the performance of Trans4EReg1 and Trans4EReg2 decreases as the dimension increases from 5 to 50. In fact, Trans4EReg1 and Trans4EReg2 with dimension 5 still outperforms all the models with dimension 50 in most of the metrics.

Table 5.3 reports the experiments with a dimension of 500. For *hasGRIDType*, Trans4E and TransE are comparable and obtain the best performances. When considering *hasTopic*, QuatE, RotatE, and Trans4E perform similarly well. Specifically, QuatE yields the best performance in Hits@1 (34.1%), while Trans4E and RotatE perform best in Hits@3 (49.2%), and Trans4E obtains the highest Hits@10 (62.8%).

Figure 5.1 and 5.2 summarize the performances of all the models for dimension 5, 50, and 500. Trans4EReg1 significantly outperforms all the models when using low dimensions and performs well also in high dimensions.

Link Prediction on Benchmark Datasets

We evaluated the performances of the Trans4E model against the competitors on a set of standard benchmark datasets with diverse relations

(N to M relations where N and M are large, N and M are small, N≫M and N≪M).

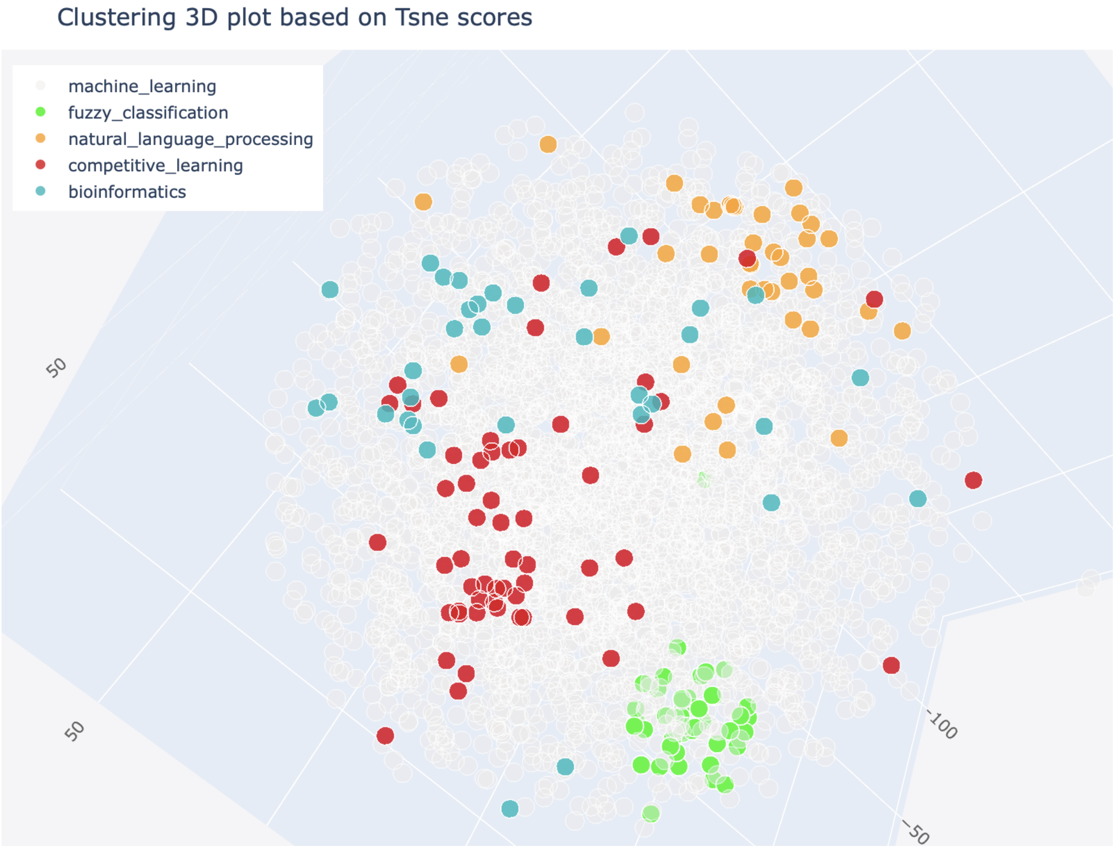


Figure 5.3: Distribution of the main topics in academia and industry.

Table 5.4 and Table 5.5 show the performances of the KGE models on the benchmark datasets FB15k, FB15k-237, WN18, and WN18RR. Trans4E outperforms the other models in Hits@3 and Hits@10 in FB15k and WN18. It also obtains a significantly better MR on FB15k-237 and WN18RR. In FB15k, the Trans4E model outperforms all the other models when considering the Hits@3 and Hits@10. In WN18, Trans4E outperforms TransE and QuatE, and obtains competitive results with respect to RotatE. To note that, these results are computed by running the models on the benchmark datasets using the best obtained hyperparameter settings where the dimension is 200, and with 20 negative samples using adversarial negative sampling [155]. The results are comparatively close in the case of FB15k-237 and WN18RR, where Trans4E has a better performance in MR.

Overall, the results show that our model outperforms other KGE models on N to M relations with N≫M and provides competitive performance on KGs with diverse relations.

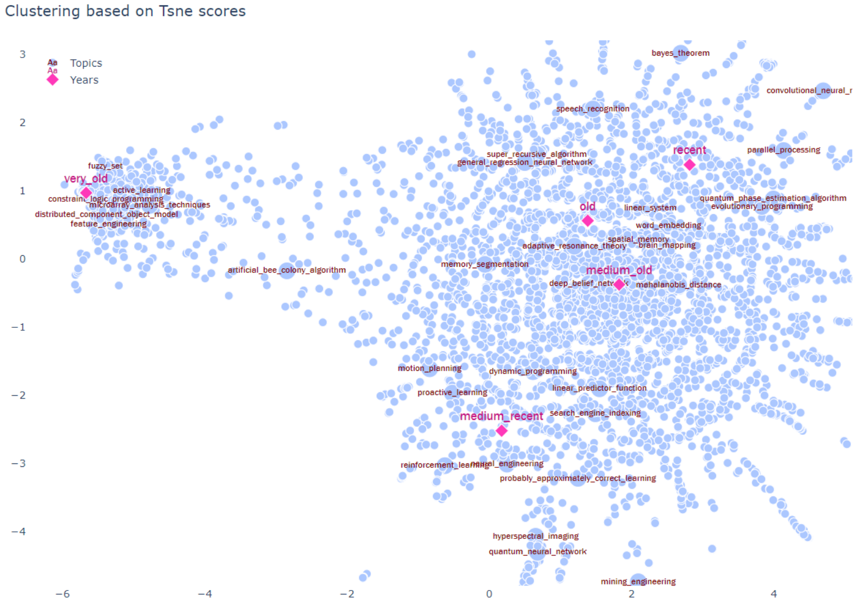


Figure 5.4: **Distribution of Topics w.r.t Years**. year >= 2015 is considered recent, year >= 2010 and year < 2015 are denoted as medium_recent, year >= 2005 and year < 2010 are medium_old, year>= 2000 and year < 2005 mean old, and anything before 2000 is very_old.

Efficiency of the Embeddings

To further investigate the representation of research topics with Trans4E, we analysed how the embeddings discriminate articles tagged with different topics.

Figure 5.3 shows the embeddings associated to the articles in AIDA+MAG in two dimensions. In order to produce it, we first selected five major topics of the machine learning venues: "fuzzy_classification", "natural_language_processing", "competitive_learning", "machine learning", and "bioinformatics". Then, we retrieved the embedding vectors of the papers tagged with those topics and visualized them by using T-SNE [92].

We can appreciate how papers with the same topics tend to cluster together. For example, papers belonging to the "fuzzy_classification" topic (green) lie within the same cluster. Note that papers in some topics such as "bioinformatics" may be associated to other topics as well (e.g. a paper may be in "bioinformatics" and use "fuzzy_classification" methods). This is why papers related to more general topics are distributed with a larger variance.

We further evaluated the ability of our model to properly distribute topics in the vector space based on their publication dates. In Figure 5.4, we illustrate the distribution of the learned vectors for the topics w.r.t their publishing years. This shows that topics such as "convolutional_neural_networks", "parallel_processing", and "speech_recognition" are correctly identified to be hot topics for the corresponding years.

The topic “word_embedding” lies in the border of recent and medium_old period indicating that even if old is still lasting.There is also a cluster of topics around the very_old time period for which the corresponding vectors are very different from the ones in other time periods. A manual analysis revealed that most of them were mostly active before the year 2000.

Chapter 6

AIDA-KG Analysis

6.1 Introduction

In this chapter, we present a scientometric analysis in which we assess whether a diverse pool of expertise within a research team can influence their scientific impact, measured as the number of citations received by the resulting research papers in the upcoming 5 years. The analysis was performed on 114,203 Computer Science papers from the Academia/Industry DynAmics (AIDA) Knowledge Graph , published within the 2010-2015 timeframe. To assess the diversity of a team, we characterise a researcher's expertise as the distribution of research topics of their paper in the previous 5 years. To this purpose, we leverage the Computer Science Ontology, which consists of 14K topics and provides a more fine-grained representation compared to the generic disciplines provided by typical scholarly datasets such as Scopus and Web of Science. We then computed the pairwise cosine similarity between each couple of authors in a paper and defined two metrics as proxy for diversity of expertise: 1) the maximum value of cosine distance between the authors, and 2) the number of connected components obtained when linking authors according to a similarity threshold.

The results show that both diversity metrics are significantly associated with the number of citations at five years. In other words, research papers authored by a research team with a wide set of skills and expertise tend to have a higher impact than the ones authored by more homogeneous teams. The remainder of the chapter is organised as follows. In paragraph 6.2 we describe the open issues, paragraph 6.3 provides an overview of the state of the art, while paragraph 6.4 outlines the materials and methodologies employed in the study. The findings are presented in Paragraph 6.5.

6.2 Open Issues

Understanding the correlation between the composition of a research team and the potential impact of their research papers is of paramount importance. Such comprehension can pave the way for the formulation of science policies and best practices aimed at propelling innovation. One commonly scrutinized aspect is the diversity of the research team across various dimensions, including but not limited to ethnicity [7], gender [110], disciplinary backgrounds [161], team size [173], among others. Notably, less attention has been dedicated to the diversity of expertise within the group of researchers.

In recent years, there has been a growing emphasis on the significance of interdisciplinary approaches and collaborative endeavors between different scientific fields. Funding agencies, scientific journals, and governmental institutions have progressively underscored the necessity of such interactions. While the current landscape may inadvertently encourage researchers to specialize narrowly within their fields, the scientific community aspires to unite its efforts to tackle societal challenges. These challenges encompass a spectrum of pressing issues, including climate change, poverty, disease, inequality, and the imperative for sustainable development. By their very nature, these challenges demand intricate and multifaceted solutions that call for the amalgamation of diverse areas of expertise. The cross-pollination of ideas from different fields of expertise can also break down the traditional barriers between disciplines and uncover unexpected insights that can drive new discoveries.

6.3 Background

In the literature, we can find a plethora of studies that analysed research team diversity across several dimensions: nationality [151], ethnicity [7, 60], institutions [74], gender [110], academic age [73], disciplinary backgrounds [161], and team size [173]. The literature consensus is that a higher diversity often leads to an increase in productivity or impact. For instance, Smith et al. (2014) showed that promoting international collaboration has important benefits for scientific visibility, quality, and impact. Likewise, research on cross-institution teams in the field of engineering, social science, and others highlighted that multi-university collaborations with top-tier universities produce high-impact papers [74]. Wu et al. (2019) analysed the size of research teams and found that small teams tend to build on less popular and potentially disruptive ideas, but also experience a citation delay, whereas larger teams work on more popular ideas and gather citations rapidly [173]. AlShebli et al. (2018) studied the effect of ethnicity, gender, academic age, and affiliations on research impact. Their analysis shows that, even if all these factors play a role, ethnicity is the most prominent one, associated with an impact gain of 10.63In this chapter, we focus instead on the diversity of expertise, which has been notoriously hard to study since scholarly datasets lack a high-quality representation of researchers' expertise.

6.4 Approach

In this paragraph, we describe the data selection and the methodologies used to assess the diversity of expertise within a paper.

6.4.1 Data Selection

To analyse whether expertise diversity is related to the number of citations, we selected 114,203 research publications fulfilling four constraints: i) they were published between 2010 and 2015, ii) they reached at least 2 citations in the following five years, iii) they were authored by at least two authors, and iv) each author had at least one publication in the five years prior the paper under analysis. We set the first constraint to compare papers in the same time period. The second condition excludes patents and other technical documents that sometimes get included in the dataset, but do not typically receive citations. The third condition is a minimum requirement to analyse the characteristics of a research team. The last condition is required to compute metrics that consider the recent expertise of the authors. In practice, we first randomly selected 150K papers from AIDA KG in the period 2010-15 (first constraint) and then removed the ones that did not meet the remaining constraints. We assessed the impact of a paper according to the number of citations received 5 years after its publication. For instance, the impact of a paper published in 2013 would be based on all the citations gathered by 2018. We then split the papers in 10 buckets of papers according to their number of citations after 5 years. Table 6.1 reports these groups alongside their frequency and the citation median.

Table 6.1: Groups of papers according to the citation ranges

Bucket Identifier	Citation ranges (c)	Citation Median	Num. of Papers
A	$2 \leq c < 5$	3	37,232
B	$5 \leq c < 10$	6	27,696
C	$10 \leq c < 15$	12	12,606
D	$15 \leq c < 20$	17	7,180
E	$20 \leq c < 30$	24	7,355
F	$30 \leq c < 40$	34	3,717
G	$40 \leq c < 50$	44	2,181
H	$50 \leq c < 100$	64	3,691
I	$100 \leq c < 150$	118	6,245
J	$c \geq 150$	226	6,292

6.4.2 Assessing Author Expertize

As a following step, we identified 363,381 authors from the 114K papers and determined their expertise. Specifically, for each author, we selected their research publications in the 5 years prior to the publication of the paper under analysis. Next, we computed the distribution of topics in these articles, i.e., we counted the times a given topic appeared in the relevant papers. We normalised this distribution over the total number of papers and subtracted the normalised topic distribution of the whole Computer Science domain. This was done to identify the topics that are more relevant to the specific author, as suggested in (Angioni et al., 2022). For instance, the final weight of the topic Machine Learning for an author will be 40

Finally, we ranked the topics based on this score and selected the top 10. We ran different experiments by testing other values between 5 and 20, but the overall results were very similar.

6.4.3 Assessing Expertise Diversity in a Team of Authors

To assess the diversity of expertise, we computed two statistical metrics on each paper. The first is the maximum value of the cosine distances computed on each couple of authors in the research team. The second metric counts the number of sub-teams having different expertise. Specifically, for a given research paper, we computed the cosine distance between each couple of authors based on their top-10 topics, generating a distribution of (N×(N-1))/2 values. The cosine distance is computed as the complement to one of the cosine similarity, and moves from 0 to 1. The higher the cosine distance the more diverse the set of topics between the two authors. In this context, the average of the cosine distance is a bad indicator, since can produce very different results for papers that a researcher would consider very similar in terms of diversity. As an example, a research team consisting of an author in Human Computer Interaction and a second author in Machine Learning may obtain a fairly high value. However, a team composed of three authors in Human-Computer Interaction and three others in Machine Learning, would produce a much lower value. Therefore, we instead used as the first diversity metric the maximum value of the distribution, which does not suffer from this issue. In order to produce a more granular metric that would reflect the different components in the team (two in the previous example), we clustered authors according to their expertise and counted the resulting number of subgroups. The higher the number of subgroups the more diverse the pool of researchers. Specifically, for each paper, we created an authorship graph G = (V,E), where V is the set of authors and E is the list of edges. We generated an edge between a pair of authors when their cosine distance is below 0.3, i.e., they have a similarity higher than 0.7, which typically indicates a high degree of similarity between two vectors. Next, we extract the number of connected components, which are the groups of authors with similar expertise. Finally, based on the number of extracted components, we characterised the paper's diversity of expertise as: i) low, with 1 or 2 components, ii) moderate, with 3 or 4 components, iii) high, with 5 or 6 components, or iv) very high, from 7 components upward. Figure 6.1 shows an example of an author network with 7 authors arranged in 3 subgroups.

6.4.4 Investigating the relationship between diversity of expertise and citations

In order to assess if the expertise diversity of the authors of a paper is significantly associated with the number of citations in the following five years, we studied the distribution of the two previously described metrics across the 10 buckets. The difference between variables was studied with the chi-square test. The correlation between distributions of continuous variables was expressed by Pearson's linear correlation coefficient r, and relative p-value. Statistical significance corresponded in both cases to $p<0.05$.

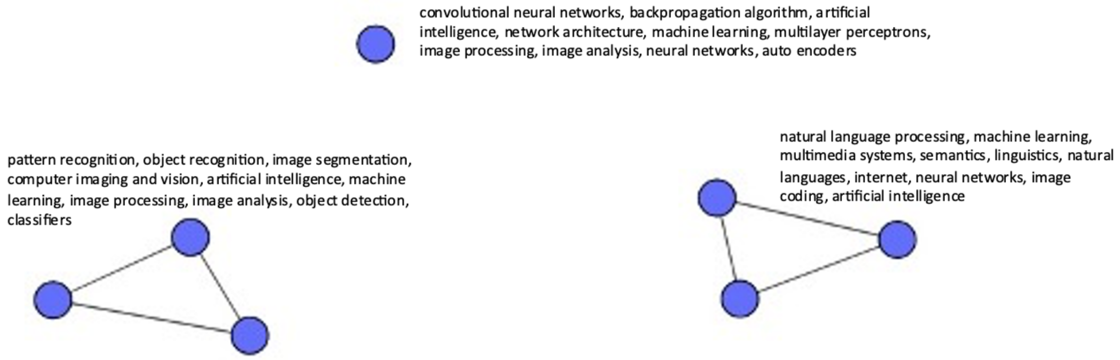

convolutional neural networks, backpropagation algorithm, artificial intelligence, network architecture, machine learning, multilayer perceptrons, image processing, image analysis, neural networks, auto encoders
pattern recognition, object recognition, image segmentation, computer imaging and vision, artificial intelligence, machine learning, image processing, image analysis, object detection, classifiers
natural language processing, machine learning, multimedia systems, semantics, linguistics, natural languages, internet, neural networks, image coding, artificial intelligence

Figure 6.1: Example of network of authors with 3 components. In this case, the paper will be characterised as having moderate diversity.

6.5 Results

In this section, we discuss our results and report relevant statistical tests.

6.5.1 Max cosine distance between authors

Figure 6.2 reports the frequency of papers for a certain maximum cosine distance over the full dataset. The most notable characteristic is the peak at 1. The specific distributions of the 10 buckets are similar in the range 0.1-0.9, but exhibit remarkable differences among them in the frequencies of 0 and 1. Therefore, we focused our analysis on these two cases. A score of 0 means that all the authors of a paper have exactly the same expertise (e.g., they all work on the very same branch of Human-Computer Interaction), while a score of 1 means that at least one author is working on completely different areas. The ratio #1/#0 can thus be used as a good indication of diversity. A higher ratio of #1/#0 will point to a higher expertise diversity.

Table 6.2 reports on the number of 0 and 1 across the buckets as well as the ratio between these two values. Figure 3 further highlights this phenomenon by showing the #1/#0 ratio against the citation median of each bucket. The Pearson correlation coefficient between the distributions of these two variables is 0.955 ($p<0.0001$), which represents a strong direct linear correlation. This seems to confirm the hypothesis that a higher expertise diversity leads to a higher number of citations, i.e., to a

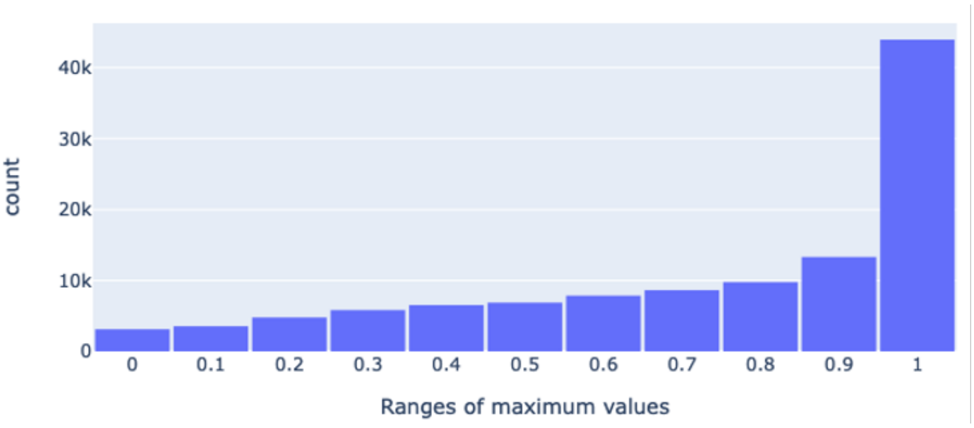


Figure 6.2: The distribution of the maximum values of cosine distance.

Identifier	# of 0s	# of 1s	#1/#0
A	1,195	14,401	12.05
B	578	10,726	18.56
C	189	4,809	25.44
D	96	2,689	28.01
E	71	2,787	39.25
F	32	1,415	44.22
G	23	820	35.65
H	28	1,398	49.93
I	33	2,406	72.91
J	25	2,351	94.04

Table 6.2: Frequency of research papers with zeros and ones according to the ranges of citations.

higher impact.

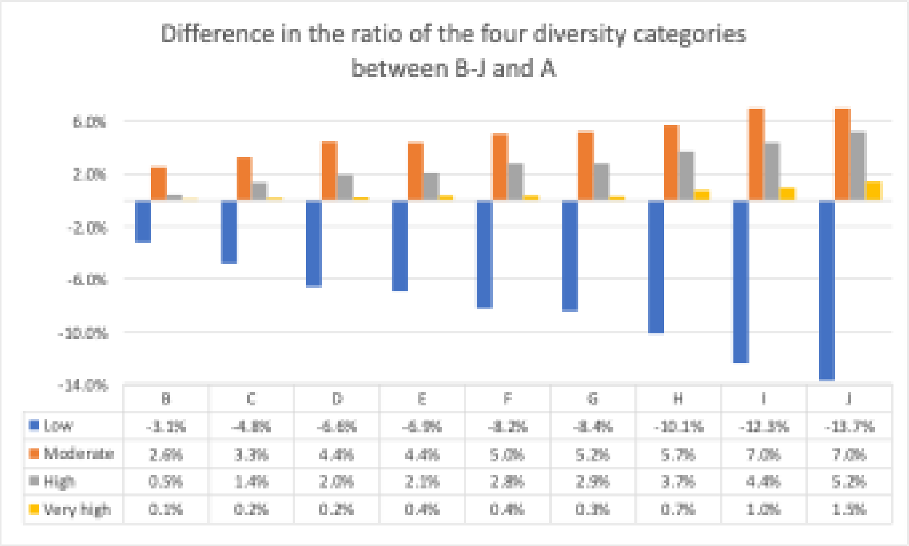
6.5.2 Number of components in the author graph

Table 6.3 reports the distribution of articles with low, moderate, high, and very high diversity as defined in Section 3.4. The percentage of papers with low diversity is inversely correlated to the median number of citations (r= -0.80, p=0.03), whereas the number of papers with high diversity is directly correlated (r=0.97, $p < 0.0001$). Hence, the ratio between the number of papers with high or very high diversity and the ones with low diversity (see last column of Table 6.3) shows a significantly high direct correlation (Pearson's r=0.90) with the median number of citations. For instance, the set of papers with less than five citations (A) includes only 3.0The chi-square test applied the distribution of the four diversity categories (low, moderate, high, and very high) between adjacent buckets i and $(i + 1)$ found a significant

	low	**moderate**	**high**	**very**	**Total**	**(very high + high)/low**
A	64.84%	32.15%	2.79%	0.23%	37,232	0.05
B	61.69%	34.71%	3.25%	0.35%	27,700	0.06
C	60.06%	35.40%	4.14%	0.40%	12,606	0.08
D	58.23%	36.56%	4.75%	0.46%	7,180	0.09
E	57.92%	36.56%	4.88%	0.64%	7,355	0.10
F	56.60%	37.18%	5.62%	0.59%	3,717	0.11
G	56.44%	37.37%	5.64%	0.55%	2,181	0.11
H	54.67%	37.83%	6.52%	0.97%	3,695	0.14
I	52.49%	39.12%	7.21%	1.18%	6,245	0.16
J	51.16%	39.16%	7.99%	1.68%	6,292	0.19
ALL	60.51%	34.91%	4.10%	0.49%	114,203	0.08

Table 6.3: Percentages of papers with low, moderate, high, and very high diversity.

difference between A and B and B and C ($p < 0.0001$), decreasing for C vs D ($p < 0.04$), and becoming not significant for the following pairs. A is also significantly different from B-J ($p < 0.0001$) and A-B is significantly different from C-J ($p < 0.0001$). Figure 6.3 further showcases this dynamic by reporting the difference in the ratio of the diversity categories between the B-J buckets and A. For instance, the difference between the ratio of high-diversity papers in J (7.99%) and A (2.78%) is 5.21%. In conclusion, the results obtained by using the four categories of diversity align with the ones based on the maximum cosine distance. In both cases, the expertise diversity metric is significantly associated with the number of citations.



	B	C	D	E	F	G	H	I	J
■ Low	-3.1%	-4.8%	-6.6%	-6.9%	-8.2%	-8.4%	-10.1%	-12.3%	-13.7%
■ Moderate	2.6%	3.3%	4.4%	4.4%	5.0%	5.2%	5.7%	7.0%	7.0%
■ High	0.5%	1.4%	2.0%	2.1%	2.8%	2.9%	3.7%	4.4%	5.2%
■ Very high	0.1%	0.2%	0.2%	0.4%	0.4%	0.3%	0.7%	1.0%	1.5%

Figure 6.3: Difference in the ratio of the diversity categories between the B-J buckets and A.

Chapter 7

Conclusion

The importance of scholarly knowledge graphs lies in their potential to revolutionize research and academia. They provide a structured, interconnected view of scholarly information, facilitating advanced search, discovery, and analysis. For the scholarly domain in this thesis we have proposed Academia/Industry DynAmics (AIDA) Knowledge graph. This resource characterizes 21 million publications and 8 million patents categorized based on research topics derived from the Computer Science Ontology. The core objective of this work was to tackle the challenge of constructing a comprehensive scholarly knowledge graph that encompasses and categorizes all dimensions of scholarly entities. It can be used to identify and analyze 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. In addition, we've shown the evaluation we performed on AIDA. Particularly, we evaluated both the pipeline for generating it and the impact of AIDA for forecasting the impact of research trends in Industry.

Within AIDA, we presented AIDA Dashboard, a tool to support the analysis and comparison of scholarly venues (conferences, journals) according to several metrics, developed within Springer Nature. We've shown that the dashboard is built on top of AIDA, and characterizes each venues according to several aspects. In fact in each venue in can assess: 1) the research area of interests, 2) analytics about every scholarly entity that publish (e.g. authors) or is useful to produce analysis (e.g. country, topics) about the venue. In addition we've discussed on the possibility and various aspects of the graphical user interface of the AIDA Dashboard and how this tool can be useful for both researchers and editors. We evaluated the dashboard both in qualitative and quantitative terms.

Another tool proposed is AIDA-Bot, a conversational agent that is built on top of AIDA Knowledge Graph, and provides accurate and factual information about the research landscape. The architecture of AIDA-Bot is based on two different modules: 1) question understanding module, and 2) response generator modules. The Question Understanding module first identifies key terms in user queries using Named-Entity Recognition, then filters out redundant terms and searches for these key terms in the AIDA Knowledge Graph. It dynamically generates grammars based on the entities found, which helps in producing relevant queries for the knowledge graph. By calculating similarity with the user's input, it selects the most suitable query for further processing by the Response Generator. While the Response Generator serves two main scenarios: for user queries that match generated queries, it runs equivalent queries on the AIDA KG to retrieve data, and for natural language responses, it employs tailored templates. We have also integrated transformers to summarize and question-answering task provided in AIDA-Bot. We've performed a user-study including open questions on the capability of the system to provide good results. The system is evaluated by the users as excellent, and the answers provided by AIDA-Bot results very good and accurate.

Subsequently, a novel embedding model for knowledge graph completion (Trans4E) is proposed. We described how this model is useful for the task of link prediction especially for knowledge graphs that have many-to-few relations, such as academic knowledge graphs. Trans4E stands for Translational Embeddings for Entity and Edge Encoding. It is based on the idea of translating entities and relations in a low-dimensional vector space, while also encoding the edge types and cardinalities. We applied the model on two large-scale academic knowledge graphs, the Academia/Industry DynAmics (AIDA) and Microsoft Academic Graph (MAG), and showed competitive results compared to other models. We've also seen how the model can help to complete the information about fields of study, affiliation types, and other attributes of academic entities.

Finally, we proposed a scientometric analysis exploring the impact of diversity in terms of the authors' expertise on the scientific impact of their papers. The approach proposes two ways of measuring the diversity of expertise: 1) The Field Diversity Index (FDI), and 2) the Field Entropy Index (FEI). We've seen how FDI measures the diversity of expertize in a team, while FEI measures the distribution of the fields in a team. In this study we found out that both FDI and FEI are positively correlated with the number of citations, meaning that papers with more diverse teams, in terms of field of expertize, tend to get more citations. It also encourages interdisciplinary collaboration and supporting diverse research communities.

List of Figures

2.1 25
2.2 Workflow for the generation of AIDA. 29
2.3 Distribution of the main topics. 34
2.4 Distribution of the topics in publications across time. 34
2.5 Distribution of the main industrial sectors. 35

3.1 AIDA Dashboard - the Overview tab of the NeurIPS conference. . . . 53
3.2 Portion of the Citation Analysis tab - The ranking of NeurIPS its focus areas. 54
3.3 Portion of the Authors tab - Authors ranked by citations in NeurIPS. 55
3.4 AIDA Dashboard - the Topics tab of the NeurIPS conference. 56
3.5 Portion of the Related Conferences tab - Conferences in Artificial Intelligence ranked by average citations. 57
3.6 Portion of the Industry tab - The main industrial sectors in NeurIPS across time. The percentage indicates the fraction of papers published in the corresponding year by companies of the underlying industrial sector. 62
3.7 The Advanced Search Panel displaying conferences in Artificial Intelligence ranked by h5-index. 63
3.8 The SUS Questionnaire results. 63
3.9 Number of votes received by each Section/Functionality. 63

4.1 Odd questions. The higher the score, the better the system. 75
4.2 Even questions. The lower the score, the better the system. 76

5.1 hasTopic for dimension 5,50 and 500 89
5.2 hasGRIDType for dimension 5,50 and 500 90
5.3 Distribution of the main topics in academia and industry. 94
5.4 **Distribution of Topics w.r.t Years**. year >= 2015 is considered recent, year >= 2010 and year < 2015 are denoted as medium_recent, year >= 2005 and year < 2010 are medium_old, year>= 2000 and year < 2005 mean old, and anything before 2000 is very_old. 95

6.1 102

6.2 The distribution of the maximum values of cosine distance. 103
6.3 Difference in the ratio of the diversity categories between the B-J buckets and A. 105

Bibliography

[1] *5th International Workshop on Chatbot Research and Design, CONVERSATIONS 2021*, vol. 13171 LNCS of *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 2022.

[2] *Expert System for Question Answering on Anomalous Events and Mitigation Strategies Using Bidirectional Transformers and Knowledge Graphs* (Abu Dhabi, 10 2022), vol. Day 3 Wed, November 02, 2022 of *Abu Dhabi International Petroleum Exhibition and Conference*. D031S084R002.

[3] Adamopoulou, E., and Moussiades, L. Chatbots: History, technology, and applications. *Machine Learning with Applications 2* (2020), 100006.

[4] Ait-Mlouk, A., and Jiang, L. Kbot: A knowledge graph based chatbot for natural language understanding over linked data. *IEEE Access 8* (2020), 149220–149230.

[5] Ali, Z., Ullah, I., Khan, A., Ullah Jan, A., and Muhammad, K. An overview and evaluation of citation recommendation models. *Scientometrics 126*, 5 (2021), 4083–4119.

[6] Alonso, R., Concas, E., and Reforgiato Recupero, D. An abstraction layer exploiting voice assistant technologies for effective human—robot interaction. *Applied Sciences 11*, 19 (2021).

[7] AlShebli, B. K., Rahwan, T., and Woon, W. L. The preeminence of ethnic diversity in scientific collaboration. *Nature communications 9*, 1 (2018), 5163.

[8] Altuntas, S., Dereli, T., and Kusiak, A. Analysis of patent documents with weighted association rules. *Technological Forecasting and Social Change 92* (2015), 249–262.

[9] Ammar, W., Groeneveld, D., Bhagavatula, C., Beltagy, I., Crawford, M., Downey, D., Dunkelberger, J., Elgohary, A., Feldman, S., Ha, V., et al. Construction of the literature graph in semantic scholar. *arXiv preprint arXiv:1805.02262* (2018).

[10] Ammari, T., Kaye, J., Tsai, J. Y., and Bentley, F. Music, search, and iot: How people (really) use voice assistants. *ACM Trans. Comput. Hum. Interact. 26*, 3 (2019), 17–1.

[11] Anderson, M. S. The complex relations between the academy and industry: Views from the literature. *The journal of higher education 72*, 2 (2001), 226–246.

[12] Anderson, M. S. The complex relations between the academy and industry: Views from the literature. *The Journal of Higher Education 72*, 2 (2001), 226–246.

[13] Angioni, S., Carta, S., Consoli, S., Reforgiato Recupero, D., and Stanciu, M. M. A big data framework based on apache spark for industry-specific lexicon generation for stock market prediction. In *The 5th International Conference on Future Networks & Distributed Systems* (2021), pp. 616–624.

[14] Angioni, S., Lincoln-DeCusatis, N., Ibba, A., and Recupero, D. R. A transformers-based approach for fine and coarse-grained classification and generation of midi songs and soundtracks. *PeerJ Computer Science 9* (2023), e1410.

[15] Angioni, S., Osborne, F., Salatino, A., Reforgiato Recupero, D., and Motta, E. Integrating knowledge graphs for comparing the scientific output of academia and industry.

[16] Angioni, S., Salatino, A., Osborne, F., Birukou, A., Recupero, D. R., and Motta, E. Assessing scientific conferences through knowledge graphs. In *International Semantic Web Conference (ISWC) 2021: Posters, Demos, and Industry Tracks* (2021), vol. 2980.

[17] Angioni, S., Salatino, A., Osborne, F., Birukou, A., Recupero, D. R., and Motta, E. Leveraging knowledge graph technologies to assess journals and conferences at springer nature. In *International Semantic Web Conference* (2022), Springer, pp. 735–752.

[18] Angioni, S., Salatino, A., Osborne, F., Recupero, D. R., and Motta, E. Aida: A knowledge graph about research dynamics in academia and industry. *Quantitative Science Studies 2*, 4 (2021), 1356–1398.

[19] Angioni, S., Salatino, A., Osborne, F., Recupero, D. R., and Motta, E. The aida dashboard: a web application for assessing and comparing scientific conferences. *IEEE Access 10* (2022), 39471–39486.

[20] Angioni, S., Salatino, A., Osborne, F., Reforgiato Recupero, D., and Motta, E. The aida dashboard: Analysing conferences with semantic technologies.

[21] Angioni, S., Salatino, A. A., Osborne, F., Recupero, D. R., and Motta, E. Integrating knowledge graphs for analysing academia and industry dynamics. In *ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium* (2020), Springer, pp. 219–225.

[22] Angioni, S., Salatino, A. A., Osborne, F., Recupero, D. R., and Motta, E. The AIDA dashboard: A web application for assessing and comparing scientific conferences. *IEEE Access 10* (2022), 39471–39486.

[23] Ankrah, S., and Omar, A.-T. Universities–industry collaboration: A systematic review. *Scandinavian Journal of Management 31*, 3 (2015), 387–408.

[24] Ankrah, S. N., Burgess, T. F., Grimshaw, P., and Shaw, N. E. Asking both university and industry actors about their engagement in knowledge transfer: What single-group studies of motives omit. *Technovation 33*, 2-3 (2013), 50–65.

[25] Athreya, R. G., Ngonga Ngomo, A.-C., and Usbeck, R. Enhancing community interactions with data-driven chatbots–the dbpedia chatbot. In *Companion Proceedings of the The Web Conference 2018* (Lyon, France, 2018), pp. 143–146.

[26] Bader, S. R., Grangel-Gonzalez, I., Nanjappa, P., Vidal, M.-E., and Maleshkova, M. A knowledge graph for industry 4.0. In *European Semantic Web Conference* (Herakleio, Greece, 2020), Springer, pp. 465–480.

[27] Bavaresco, R., Silveira, D., Reis, E., Barbosa, J., Righi, R., Costa, C., Antunes, R., Gomes, M., Gatti, C., Vanzin, M., et al. Conversational agents in business: A systematic literature review and future research directions. *Computer Science Review 36* (2020), 100239.

[28] Beck, M., Rizvi, S. T. R., Dengel, A., and Ahmed, S. From automatic keyword detection to ontology-based topic modeling. In *International Workshop on Document Analysis Systems* (2020), Springer, pp. 451–465.

[29] Belleau, F., Nolin, M.-A., Tourigny, N., Rigault, P., and Morissette, J. Bio2rdf: towards a mashup to build bioinformatics knowledge systems. *Journal of biomedical informatics 41*, 5 (2008), 706–716.

[30] Bellini, V., Biancofiore, G. M., Di Noia, T., Di Sciascio, E., Narducci, F., and Pomo, C. Guapp: A conversational agent for job recommendation for the italian public administration. In *Proceedings of the IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS 2020)* (2020).

[31] Bharti, U., Bajaj, D., Batra, H., Lalit, S., Lalit, S., and Gangwani, A. Medbot: Conversational artificial intelligence powered chatbot for delivering tele-health after covid-19. In *2020 5th International Conference on Communication and Electronics Systems (ICCES)* (Budva, Montenegro, 2020), pp. 870–875.

[32] Bikard, M., Vakili, K., and Teodoridis, F. When collaboration bridges institutions: The impact of university–industry collaboration on academic productivity. *Organization Science 30*, 2 (2019), 426–445.

[33] Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *J. Mach. Learn. Res. 3*, null (Mar. 2003), 993–1022.

[34] Bockhorst, J., Conathan, D., and Fung, G. M. Knowledge graph-driven conversational agents. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)* (Vancouver, Canada, 2019).

[35] Borah, B., Pathak, D., Sarmah, P., Som, B., and Nandi, S. Survey of textbased chatbot in perspective of recent technologies. In *Computational Intelligence, Communications, and Business Analytics* (Singapore, 2019), J. K. Mandal, S. Mukhopadhyay, P. Dutta, and K. Dasgupta, Eds., Springer Singapore, pp. 84–96.

[36] Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O. Translating embeddings for modeling multi-relational data. In *Advances in NIPS* (2013).

[37] Borges, M. V. M., and dos Reis, J. C. Semantic-enhanced recommendation of video lectures. In *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)* (2019), vol. 2161, IEEE, pp. 42–46.

[38] Brixey, J., Hoegen, R., Lan, W., Rusow, J., Singla, K., Yin, X., Artstein, R., and Leuski, A. Shihbot: A facebook chatbot for sexual health information on hiv/aids. In *Proceedings of the 18th annual SIGdial meeting on discourse and dialogue* (Saarbrücken, Germany, 2017), pp. 370–373.

[39] Brooke, J. Sus: A 'quick and dirty' usability scale. *Usability evaluation in industry 189* (1996).

[40] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D.

Language models are few-shot learners. In *Advances in Neural Information Processing Systems* (2020), H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., pp. 1877–1901.

[41] Buscaldi, D., Dessì, D., Motta, E., Osborne, F., and Reforgiato Recupero, D. Mining scholarly publications for scientific knowledge graph construction. In *Presented at ESWC 16th* (2019).

[42] Callejas, Z., and Griol, D. Conversational agents for mental health and wellbeing. In *Dialog Systems: A Perspective from Language, Logic and Computation*, T. Lopez-Soto, Ed. Springer International Publishing, Cham, 2021, pp. 219–244.

[43] Chaoua, I., Recupero, D., Consoli, S., Härmä, A., and Helaoui, R. Detecting and tracking ongoing topics in psychotherapeutic conversations. In *1st Joint Workshop on AI in Health, AIH 2018* (Stockholm, Sweden, 2018), vol. 2142, pp. 97–108.

[44] Chatzopoulos, S., Vergoulis, T., Kanellos, I., Dalamagas, T., and Tryfonopoulos, C. Artsim: improved estimation of current impact for recent articles. In *ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium* (2020), Springer, pp. 323–334.

[45] Chatzopoulos, S., Vergoulis, T., Kanellos, I., Dalamagas, T., and Tryfonopoulos, C. Artsim: Improved estimation of current impact for recent articles. In *ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium* (Cham, 2020), L. Bellatreche, M. Bieliková, O. Boussaïd, B. Catania, J. Darmont, E. Demidova, F. Duchateau, M. Hall, T. Merčun, B. Novikov, C. Papatheodorou, T. Risse, O. Romero, L. Sautot, G. Talens, R. Wrembel, and M. Žumer, Eds., Springer International Publishing, pp. 323–334.

[46] Chen, H., Liu, X., Yin, D., and Tang, J. A survey on dialogue systems: Recent advances and new frontiers. *SIGKDD Explor. Newsl. 19*, 2 (Nov. 2017), 25–35.

[47] Chicaiza, J., and Reátegui, R. Using domain ontologies for text classification. a use case to classify computer science papers. In *Iberoamerican Knowledge Graphs and Semantic Web Conference* (2020), Springer, pp. 166–180.

[48] Choi, S., and Jun, S. Vacant technology forecasting using new bayesian patent clustering. *Technology Analysis & Strategic Management 26*, 3 (2014), 241–251.

[49] Chung, P., and Sohn, S. Y. Early detection of valuable patents using a deep learning model: Case of semiconductor industry. *Technological Forecasting and Social Change 158* (2020), 120146.

[50] Costa, J. P., Rei, L., Stopar, L., Fuart, F., Grobelnik, M., Mladenic, D., Novalija, I., Staines, A., Paakkonen, J., Konttila, J., Bidaurrazaga, J., Belar, O., Henderson, C., Epelde, G., Gabarain, M. A., Carlin, P., and Wallace, J. Newsmesh: A new classifier designed to annotate health news with mesh headings. *Artificial Intelligence in Medicine 114* (2021), 102053.

[51] Cui, L., Huang, S., Wei, F., Tan, C., Duan, C., and Zhou, M. Superagent: A customer service chatbot for e-commerce websites. In *Proceedings of ACL 2017, System Demonstrations* (Vancouver, Canada, 2017), Microsoft Research, pp. 97–102.

[52] Cui, L., Huang, S., Wei, F., Tan, C., Duan, C., and Zhou, M. SuperAgent: A customer service chatbot for E-commerce websites. In *Proceedings of ACL 2017, System Demonstrations* (Vancouver, Canada, July 2017), Association for Computational Linguistics, pp. 97–102.

[53] Dessì, D., Osborne, F., Recupero, D. R., Buscaldi, D., and Motta, E. Generating knowledge graphs by employing natural language processing and machine learning techniques within the scholarly domain. *Future Generation Computer Systems 116* (2021), 253–264.

[54] Divya, S., Indumathi, V., Ishwarya, S., Priyasankari, M., and Devi, S. K. A self-diagnosis medical chatbot using artificial intelligence. *Journal of Web Development and Web Designing 3*, 1 (2018), 1–7.

[55] Dörpinghaus, J., and Jacobs, M. Knowledge detection and discovery using semantic graph embeddings on large knowledge graphs generated on text mining results. In *2020 15th Conference on Computer Science and Information Systems (FedCSIS)* (2020), pp. 169–178.

[56] Elbasri, H., Haddi, A., and Allali, H. Improving e-learning by integrating a metacognitive agent, Oct 2018.

[57] Fathalla, S., Auer, S., and Lange, C. Towards the semantic formalization of science. In *Proc. of 35th Annual ACM Symposium on Applied Comp.* (2020), pp. 2057–2059.

[58] Fensel, D., Şimşek, U., Angele, K., Huaman, E., Kärle, E., Panasiuk, O., Toma, I., Umbrich, J., and Wahler, A. *Knowledge Graphs Methodology, Tools and Selected Use Cases.* 2020.

[59] Franceschet, M. The role of conference publications in cs. *Communications of the ACM 53*, 12 (2010), 129–132.

[60] Freeman, R. B., and Huang, W. Collaborating with people like me: Ethnic coauthorship within the united states. *Journal of Labor Economics 33*, S1 (2015), S289–S318.

[61] Goyal, P., and Ferrara, E. Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems 151* (2018), 78–94.

[62] Grimpe, C., and Hussinger, K. Formal and informal knowledge and technology transfer from academia to industry: Complementarity effects and innovation performance. *Industry and Innovation 20*, 8 (2013), 683–700.

[63] Groth, P., Gibson, A., and Velterop, J. The anatomy of a nanopublication. *Information Services & Use 30*, 1-2 (2010), 51–56.

[64] Guilarte, O. F., Barbosa, S. D. J., and Pesco, S. RelPath: an interactive tool to visualize branches of studies and quantify the expertise of authors by citation paths. *Scientometrics 126*, 6 (2021), 4871–4897.

[65] Henk, V., Vahdati, S., Nayyeri, M., Ali, M., Yazdi, H. S., and Lehmann, J. Metaresearch recommendations using knowledge graph embeddings. In *RecNLP workshop of AAAI Conference* (2019).

[66] Höffner, K., Walter, S., Marx, E., Usbeck, R., Lehmann, J., and Ngonga Ngomo, A.-C. Survey on challenges of question answering in the semantic web. *Semantic Web 8*, 6 (2017), 895–920.

[67] Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., Melo, G. D., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., et al. Knowledge graphs. *ACM Computing Surveys (CSUR) 54*, 4 (2021), 1–37.

[68] Hsu, P., Zhao, J., Liao, K., Liu, T., and Wang, C. Allergybot: A chatbot technology intervention for young adults with food allergies dining out. In *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems* (Denver, Colorado, 2017), pp. 74–79.

[69] Huang, M.-H., Yang, H.-W., and Chen, D.-Z. Industry–academia collaboration in fuel cells: A perspective from paper and patent analysis. *Scientometrics 105*, 2 (2015), 1301–1318.

[70] Hussain, S., Sianaki, O. A., and Ababneh, N. A survey on conversational agents/chatbots classification and design techniques. In *Workshops of the International Conference on Advanced Information Networking and Applications* (Matsue, Japan, 2019), Springer, pp. 946–956.

[71] Jaradeh, M. Y., Oelen, A., Farfar, K. E., Prinz, M., D'Souza, J., Kismihók, G., Stocker, M., and Auer, S. Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge. In *Proceedings of the 10th International Conference on Knowledge Capture* (2019), pp. 243–246.

[72] Jefferson, O. A., Koellhofer, D., Warren, B., and Jefferson, R. The lens metarecord and lensid: An open identifier system for aggregated metadata and versioning of knowledge artefacts, Nov 2019.

[73] Jones, B. F., and Weinberg, B. A. Age dynamics in scientific creativity. *Proceedings of the national academy of sciences 108*, 47 (2011), 18910–18914.

[74] Jones, B. F., Wuchty, S., and Uzzi, B. Multi-university research teams: Shifting impact, geography, and stratification in science. *science 322*, 5905 (2008), 1259–1262.

[75] Jose, V., Jagathy Raj, V. P., and George, S. K. Ontology-based information extraction framework for academic knowledge repository. In *Proceedings of Fifth International Congress on Information and Communication Technology* (Singapore, 2021), X.-S. Yang, S. Sherratt, N. Dey, and A. Joshi, Eds., Springer Singapore, pp. 73–80.

[76] Karatas, N., Tamura, S., Fushiki, M., and Okada, M. Multi-party conversation of driving agents: The effects of overhearing information on lifelikeness and distraction. In *Proceedings of the 6th International Conference on Human-Agent Interaction* (New York, NY, USA, 2018), HAI '18, Association for Computing Machinery, p. 84–91.

[77] Khalid, S., Wu, S., Wahid, A., Alam, A., and Ullah, I. An effective scholarly search by combining inverted indices and structured search with citation networks analysis. *IEEE Access 9* (2021), 120210–120226.

[78] Knoth, P., and Zdrahal, Z. Core: connecting repositories in the open access domain. In *CERN Workshop on Innovations in Scholarly Communication (OAI7)* (2011). Poster Session ID: 53.

[79] Knoth, P., and Zdrahal, Z. Core: three access levels to underpin open access. *D-Lib Magazine 18*, 11/12 (2012), 1–13.

[80] Kuhn, T., Chichester, C., Krauthammer, M., Queralt-Rosinach, N., Verborgh, R., Giannakopoulos, G., Ngomo, A.-C. N., Viglianti, R., and Dumontier, M. Decentralized provenance-aware publishing with nanopublications. *PeerJ C. S. 2* (2016), e78.

[81] KUHN, T. S. The structure of scientific revolutions: University of chicago press. *Original edition* (1962).

[82] LANDIS, J. R., AND KOCH, G. G. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.

[83] LARANJO, L., DUNN, A. G., TONG, H. L., KOCABALLI, A. B., CHEN, J., BASHIR, R., SURIAN, D., GALLEGO, B., MAGRABI, F., LAU, A. Y., ET AL. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association 25*, 9 (2018), 1248–1258.

[84] LARIVIÈRE, V., MACALUSO, B., MONGEON, P., SILER, K., AND SUGIMOTO, C. R. Vanishing industries and the rising monopoly of universities in published research. *PLOS ONE 13* (08 2018), 1–10.

[85] LI, L., LEE, K. Y., EMOKPAE, E., AND YANG, S.-B. What makes you continuously use chatbot services? evidence from chinese online travel agencies. *Electronic Markets* (2021).

[86] LI, L., WANG, P., WANG, Y., JIANG, J., TANG, B., YAN, J., WANG, S., AND LIU, Y. Prtransh: Embedding probabilistic medical knowledge from real world emr data. *arXiv preprint arXiv:1909.00672* (2019).

[87] LI, X., CHEN, Y.-N., LI, L., GAO, J., AND CELIKYILMAZ, A. End-to-end task-completion neural dialogue systems. In *IJCNLP* (Taiwan, 2017).

[88] LIU, Y.-H., ARNOLD, A., DUPONT, G., KOBUS, C., AND LANCELOT, F. Evaluation of conversational agents for aerospace domain. In *Proceedings of the Joint Conference of the Information Retrieval Communities in Europe (CIRCLE 2020)* (07 2020).

[89] LÖFFLER, F., WESP, V., BABALOU, S., KAHN, P., LACHMANN, R., SATELI, B., WITTE, R., AND KÖNIG-RIES, B. Scholarlensviz: A visualization framework for transparency in semantic user profiles. In *Proceedings of the ISWC 2020 Demos and Industry Tracks: From Novel Ideas to Industrial Practice co-located with 19th International Semantic Web Conference (ISWC 2020), Globally online, November 1-6, 2020 (UTC).* (2020), K. Taylor, R. Gonçalves, F. Lecue, and J. Yan, Eds.

[90] LOMMATZSCH, A. A next generation chatbot-framework for the public administration. In *Innovations for Community Services* (Zilina, Slovenia, 2018), M. Hodoň, G. Eichler, C. Erfurth, and G. Fahrnberger, Eds., Springer International Publishing, pp. 127–141.

[91] LULA, P., DOSPINESCU, O., HOMOCIANU, D., AND SIRETEANU, N.-A. An advanced analysis of cloud computing concepts based on the computer science ontology. *Computers, Materials & Continua 66*, 3 (2021), 2425–2443.

[92] Maaten, L. v. d., and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research 9*, Nov (2008), 2579–2605.

[93] Mannocci, A., Osborne, F., and Motta, E. Geographical trends in academic conferences: An analysis of authors' affiliations. *Data Science 2*, 1-2 (2019), 181–203.

[94] Marinakis, Y. D. Forecasting technology diffusion with the richards model. *Technological Forecasting and Social Change 79*, 1 (2012), 172–179.

[95] Meloni, A., Angioni, S., Salatino, A., Osborne, F., Recupero, D. R., and Motta, E. Integrating conversational agents and knowledge graphs within the scholarly domain. *IEEE Access 11* (2023), 22468–22489.

[96] Meloni, A., Angioni, S., Salatino, A., Osborne, F., Reforgiato Recupero, D., and Motta, E. Aida-bot: A conversational agent to explore scholarly knowledge graphs.

[97] Michaudel, Q., Ishihara, Y., and Baran, P. S. Academia–industry symbiosis in organic chemistry. *Accounts of Chemical Research 48*, 3 (2015), 712–721. PMID: 25702529.

[98] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2* (USA, 2013), NIPS'13, Curran Associates Inc., pp. 3111–3119.

[99] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (2013), pp. 3111–3119.

[100] Mohan, S., and Chowdhary, C. An ai-based chatbot using deep learning. In *Intelligent Systems: Advances in Biometric Systems, Soft Computing, Image Processing, and Data Analytics*. Apple Academic Press, London, UK, 12 2019, ch. 12, pp. 231–242.

[101] Montenegro, J. L. Z., da Costa, C. A., and da Rosa Righi, R. Survey of conversational agents in health. *Expert Systems with Applications 129* (2019), 56–67.

[102] Mora-Cantallops, M., Sánchez-Alonso, S., and García-Barriocanal, E. A systematic literature review on wikidata. *Data Technologies and Applications* (2019).

[103] Nayyeri, M., Cil, G. M., Vahdati, S., Osborne, F., Rahman, M., Angioni, S., Salatino, A., Recupero, D. R., Vassilyeva, N., Motta, E., et al. Trans4e: Link prediction on scholarly knowledge graphs. *Neurocomputing 461* (2021), 530–542.

[104] Nayyeri, M., Gil, G., Vahdati, S., Osborne, F., Kravchenko, A., Angioni, S., Salatino, A., Recupero, D., Motta, E., and Lehmann, J. Link prediction using numerical weights for knowledge graph completion within the scholarly domain. In *Proc. of ESWC* (2021), vol. 21.

[105] Nayyeri, M., Vahdati, S., Zhou, X., Yazdi, H. S., and Lehmann, J. Embedding-based recommendations on scholarly knowledge graphs. In *European Semantic Web Conference* (2020), Springer, pp. 255–270.

[106] Nayyeri, M., Xu, C., Vahdati, S., Vassilyeva, N., Sallinger, E., Yazdi, H. S., and Lehmann, J. Fantastic knowledge graph embeddings and how to find the right space for them. In *International Semantic Web Conference* (2020), Springer, pp. 438–455.

[107] Nayyeri, M., Xu, C., Vahdati, S., Vassilyeva, N., Sallinger, E., Yazdi, H. S., and Lehmann, J. Fantastic knowledge graph embeddings and how to find the right space for them. In *ISWC* (2020).

[108] Ni, L., Lu, C., Liu, N., and Liu, J. Mandy: Towards a smart primary care chatbot application. In *International symposium on knowledge and systems sciences* (Bangkok, Thailand, 2017), Springer, pp. 38–52.

[109] Nickel, M., Murphy, K., Tresp, V., and Gabrilovich, E. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE 104*, 1 (2016).

[110] Nielsen, F. Å., Mietchen, D., and Willighagen, E. Scholia, scientometrics and wikidata. In *The Semantic Web: ESWC 2017 Satellite Events* (Cham, 2017), E. Blomqvist, K. Hose, H. Paulheim, A. Ławrynowicz, F. Ciravegna, and O. Hartig, Eds., Springer International Publishing, pp. 237–259.

[111] Nuzzolese, A. G., Gentile, A. L., Presutti, V., and Gangemi, A. Semantic web conference ontology-a refactoring solution. In *European Semantic Web Conference* (2016), Springer, pp. 84–87.

[112] Oh, K.-J., Lee, D., Ko, B., and Choi, H.-J. A chatbot for psychiatric counseling in mental healthcare service based on emotional dialogue analysis and sentence generation. In *2017 18th IEEE International Conference on Mobile Data Management (MDM)* (KAIST, Daejeon, 2017), IEEE, pp. 371–375.

[113] Okonkwo, C. W., and Ade-Ibijola, A. Chatbots applications in education: A systematic review. *Computers and Education: Artificial Intelligence 2* (2021), 100033.

[114] OpenAI. Gpt-4 technical report, 2023.

[115] Osborne, F., Mannocci, A., and Motta, E. Forecasting the spreading of technologies in research communities. In *Proceedings of the Knowledge Capture Conference* (New York, NY, USA, 2017), K-CAP 2017, ACM, pp. 1:1–1:8.

[116] Osborne, F., and Motta, E. Klink-2: integrating multiple web sources to generate semantic topic networks. In *ISWC* (2015), Springer, pp. 408–424.

[117] Osborne, F., and Motta, E. Pragmatic ontology evolution: reconciling user requirements and application performance. In *International Semantic Web Conference* (Monterey, USA, 2018), Springer, pp. 495–512.

[118] Osborne, F., Muccini, H., Lago, P., and Motta, E. Reducing the effort for systematic reviews in software engineering. *Data Science 2*, 1-2 (2019), 311–340.

[119] Osborne, F., Salatino, A., Birukou, A., and Motta, E. Automatic classification of springer nature proceedings with smart topic miner. In *The Semantic Web – ISWC 2016* (Cham, 2016), P. Groth, E. Simperl, A. Gray, M. Sabou, M. Krötzsch, F. Lecue, F. Flöck, and Y. Gil, Eds., Springer Int. Publishing, pp. 383–399.

[120] Pang, R. Y., Parrish, A., Joshi, N., Nangia, N., Phang, J., Chen, A., Padmakumar, V., Ma, J., Thompson, J., He, H., and Bowman, S. QuALITY: Question answering with long input texts, yes! In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Seattle, United States, July 2022), Association for Computational Linguistics, pp. 5336–5358.

[121] Paulheim, H. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic web 8*, 3 (2017), 489–508.

[122] Peroni, S., and Shotton, D. The spar ontologies. In *International Semantic Web Conference* (2018), Springer, pp. 119–136.

[123] Peroni, S., and Shotton, D. Opencitations, an infrastructure organization for open scholarship. *Quantitative Science Studies 1*, 1 (2020), 428–444.

[124] Powell, W. W., and Snellman, K. The knowledge economy. *Annual Review of Sociology 30*, 1 (2004), 199–220.

[125] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog 1*, 8 (2019), 9.

[126] Rahdari, B., Brusilovsky, P., and Javadian Sabet, A. *Connecting Students with Research Advisors Through User-Controlled Recommendation.* Association for Computing Machinery, New York, NY, USA, 2021, p. 745–748.

[127] Ramadhan, M. H., Malik, V. I., and Sjafrizal, T. Artificial neural network approach for technology life cycle construction on patent data. In *2018 5th International Conference on Industrial Engineering and Applications (ICIEA)* (2018), IEEE, pp. 499–503.

[128] Ramesh, K., Ravishankaran, S., Joshi, A., and Chandrasekaran, K. A survey of design techniques for conversational agents. In *International conference on information, communication and computing technology* (New Delhi, India, 2017), Springer, pp. 336–350.

[129] Ranoliya, B. R., Raghuwanshi, N., and Singh, S. Chatbot for university related faqs. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (Manipal, India, 2017), IEEE, pp. 1525–1530.

[130] Rastogi, A., Zang, X., Sunkara, S., Gupta, R., and Khaitan, P. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence* (New York, USA, 2020), vol. 34, pp. 8689–8696.

[131] Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Shuster, K., Smith, E. M., Boureau, Y.-L., and Weston, J. Recipes for building an open-domain chatbot. In *EACL* (Online, 2021).

[132] Rooein, D., Bianchini, D., Leotta, F., Mecella, M., Paolini, P., and Pernici, B. achat-wf: Generating conversational agents for teaching business process models. *Software and Systems Modeling* (2021).

[133] Rossanez, A., Reis, J., and Torres, R. D. S. Representing scientific literature evolution via temporal knowledge graphs. CEUR Workshop Proceedings.

[134] S, A., Karthik, N., and K S, S. Comparative study on voice based chat bots. *International Journal of Computer Sciences and Engineering 6* (12 2018), 172–175.

[135] Salatino, A., Angioni, S., Osborne, F., Recupero, D. R., and Motta, E. Diversity of expertise is key to scientific impact: a large-scale analysis in the field of computer science. *arXiv preprint arXiv:2306.15344* (2023).

[136] Salatino, A., Osborne, F., and Motta, E. Researchflow: Understanding the knowledge flow between academia and industry. In *Knowledge Engineering and Knowledge Management – 22nd International Conference, EKAW 2020* (2020).

[137] Salatino, A. A., Osborne, F., Birukou, A., and Motta, E. Improving editorial workflow and metadata quality at springer nature. In *The Semantic Web – ISWC 2019* (Cham, 2019), C. Ghidini, O. Hartig, M. Maleshkova, V. Svátek, I. Cruz, A. Hogan, J. Song, M. Lefrançois, and F. Gandon, Eds., Springer Int. Publishing, pp. 507–525.

[138] Salatino, A. A., Osborne, F., Thanapalasingam, T., and Motta, E. The cso classifier: Ontology-driven detection of research topics in scholarly articles. In *Digital Libraries for Open Knowledge* (Cham, 2019), A. Doucet, A. Isaac, K. Golub, T. Aalberg, and A. Jatowt, Eds., Springer International Publishing, pp. 296–311.

[139] Salatino, A. A., Thanapalasingam, T., and Mannocci, A. angelosalatino/cso-classifier: CSO Classifier v2.3.2, Aug. 2019.

[140] Salatino, A. A., Thanapalasingam, T., Mannocci, A., Birukou, A., Osborne, F., and Motta, E. The computer science ontology: A comprehensive automatically-generated taxonomy of research areas. *Data Intelligence 2*, 3 (2020), 379–416.

[141] Salatino, A. A., Thanapalasingam, T., Mannocci, A., Osborne, F., and Motta, E. Classifying research papers with the computer science ontology. In *ISWC (P&D/Industry/BlueSky). CEUR Workshop Proceedings* (2018), vol. 2180.

[142] Salatino, A. A., Thanapalasingam, T., Mannocci, A., Osborne, F., and Motta, E. The computer science ontology: A large-scale taxonomy of research areas. In *The Semantic Web – ISWC 2018* (Cham, 2018), D. Vrandečić, K. Bontcheva, M. C. Suárez-Figueroa, V. Presutti, I. Celino, M. Sabou, L.-A. Kaffee, and E. Simperl, Eds., Springer Int. Publishing, pp. 187–205.

[143] Sarosa, M., Kusumawardani, M., Suyono, A., and Wijaya, M. H. Developing a social media-based chatbot for english learning. *IOP Conference Series: Materials Science and Engineering 732*, 1 (jan 2020), 012074.

[144] Satopaa, V., Albrecht, J., Irwin, D., and Raghavan, B. Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st International Conference on Distributed Computing Systems Workshops* (June 2011), pp. 166–171.

[145] Schneider, J., Ciccarese, P., Clark, T., and Boyce, R. D. Using the micropublications ontology and the open annotation data model to represent evidence within a drug-drug interaction knowledge base. In *Workshop on Linked Science 2014—Making Sense Out of Data (LISC2014) at ISWC 2014* (2014).

[146] Schwartz, D. L., and Sichelman, T. Data sources on patents, copyrights, trademarks, and other intellectual property. In *Research Handbook on the Economics of Intellectual Property Law*. Edward Elgar Publishing, 2019.

[147] Shah, P., Hakkani-Tür, D., Liu, B., and Tür, G. Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)* (New Orleans - Louisiana, June 2018), Association for Computational Linguistics, pp. 41–51.

[148] Shotton, D. Semantic publishing: the coming revolution in scientific journal publishing. *Learned Publishing 22*, 2 (2009), 85–94.

[149] Singh, J., Joesph, M. H., and Jabbar, K. B. A. Rule-based chabot for student enquiries. *Journal of Physics: Conference Series 1228*, 1 (may 2019), 012060.

[150] Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-J., and Wang, K. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web* (2015), pp. 243–246.

[151] Smith, M. J., Weinberger, C., Bruna, E. M., and Allesina, S. The scientific impact of nations: Journal placement and citation performance. *PloS one 9*, 10 (2014), e109195.

[152] Stanovsky, G., Gruhl, D., and Mendes, P. Recognizing mentions of adverse drug reaction in social media using knowledge-infused recurrent models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers* (2017), pp. 142–151.

[153] Stasaski, K., and Hearst, M. Semantic diversity in dialogue with natural language inference. In *Proceedings of the 2022 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies (Seattle, United States, July 2022), Association for Computational Linguistics, pp. 85–98.

[154] Stilgoe, J. Who's driving innovation? *New Technologies and the Collaborative State. Cham, Switzerland: Palgrave Macmillan* (2020).

[155] Sun, Z., Deng, Z.-H., Nie, J.-Y., and Tang, J. Rotate: Knowledge graph embedding by relational rotation in complex space. In *International Conference on Learning Representations* (2019).

[156] Thanapalasingam, T., Osborne, F., Birukou, A., and Motta, E. Ontology-based recommendation of editorial products. In *The Semantic Web – ISWC 2018* (Cham, 2018), D. Vrandečić, K. Bontcheva, M. C. Suárez-Figueroa, V. Presutti, I. Celino, M. Sabou, L.-A. Kaffee, and E. Simperl, Eds., Springer Int. Publishing, pp. 341–358.

[157] Thongyoo, P., Anantapanya, P., Jamsri, P., and Chotipant, S. A personalized food recommendation chatbot system for diabetes patients. In *Cooperative Design, Visualization, and Engineering* (Bangkok, Thailand, 2020), Y. Luo, Ed., Springer International Publishing, pp. 19–28.

[158] Tosi, M. D. L., and dos Reis, J. C. Scikgraph: A knowledge graph approach to structure a scientific field. *Journal of Informetrics 15*, 1 (2021), 101109.

[159] Tran, H. N., and Takasu, A. Exploring scholarly data by semantic query on knowledge graph embedding space. In *International Conference on Theory and Practice of Digital Libraries* (2019), Springer, pp. 154–162.

[160] Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., and Bouchard, G. Complex embeddings for simple link prediction. In *International Conference on Machine Learning* (2016), pp. 2071–2080.

[161] Uzzi, B., Mukherjee, S., Stringer, M., and Jones, B. Atypical combinations and scientific impact. *Science 342*, 6157 (2013), 468–472.

[162] Vaira, L., Bochicchio, M. A., Conte, M., Casaluci, F. M., and Melpignano, A. Mamabot: a system based on ml and nlp for supporting women and families during pregnancy. In *Proceedings of the 22nd International Database Engineering & Applications Symposium* (Villa San Giovanni, Italy, 2018), pp. 273–277.

[163] van Eck, N. J., and Waltman, L. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics 84*, 2 (2010), 523–538.

[164] van Noordt, C., and Misuraca, G. New wine in old bottles: Chatbots in government - exploring the transformative impact of chatbots in public service delivery. In *ePart* (2019), pp. 49–59.

[165] Vergoulis, T., Chatzopoulos, S., Dalamagas, T., and Tryfonopoulos, C. Veto: Expert set expansion in academia. In *Digital Libraries for Open Knowledge* (Cham, 2020), M. Hall, T. Merčun, T. Risse, and F. Duchateau, Eds., Springer International Publishing, pp. 48–61.

[166] Vergoulis, T., Chatzopoulos, S., Dalamagas, T., and Tryfonopoulos, C. Veto: Expert set expansion in academia. In *Digital Libraries for Open Knowledge* (Cham, 2020), M. Hall, T. Merčun, T. Risse, and F. Duchateau, Eds., Springer International Publishing, pp. 48–61.

[167] Wang, K., Shen, Z., Huang, C., Wu, C.-H., Dong, Y., and Kanakia, A. Microsoft academic graph: When experts are not enough. *Quantitative Science Studies 1*, 1 (2020), 396–413.

[168] Wang, Q., Mao, Z., Wang, B., and Guo, L. Knowledge graph embedding: A survey of approaches and applications. *IEEE TKDE 29*, 12 (2017).

[169] Wang, W., Liu, J., Tang, T., Tuarob, S., Xia, F., Gong, Z., and King, I. Attributed collaboration network embedding for academic relationship mining. *ACM Transactions on the Web (TWEB) 15*, 1 (2020), 1–20.

[170] Weber, F., Wambsganss, T., Rüttimann, D., and Söllner, M. Pedagogical agents for interactive learning: A taxonomy of conversational agents in education. In *ICIS 2021 Proceedings* (2021).

[171] Weinstein, L. B., Kellar, G. M., and Hall, D. C. Comparing topic importance perceptions of industry and business school faculty: Is the tail wagging the dog? *Academy of Educational Leadership Journal 20*, 2 (2016), 62.

[172] Wolstencroft, K., Haines, R., Fellows, D., Williams, A., Withers, D., Owen, S., Soiland-Reyes, S., Dunlop, I., Nenadic, A., Fisher, P., et al. The taverna workflow suite: designing and executing workflows of web services on the desktop, web or in the cloud. *Nucleic acids research 41*, W1 (2013), W557–W561.

[173] Wu, L., Wang, D., and Evans, J. A. Large teams develop and small teams disrupt science and technology. *Nature 566*, 7744 (2019), 378–382.

[174] Xu, A., Liu, Z., Guo, Y., Sinha, V., and Akkiraju, R. A new chatbot for customer service on social media. In *Proceedings of the 2017 CHI conference on human factors in computing systems* (Denver, Colorado, 2017), pp. 3506–3510.

[175] Yao, L., Zhang, Y., Wei, B., Jin, Z., Zhang, R., Zhang, Y., and Chen, Q. Incorporating knowledge graph embeddings into topic modeling. In *Thirty-First AAAI Conference on Artificial Intelligence* (2017).

[176] Zang, X., and Niu, Y. The forecast model of patents granted in colleges based on genetic neural network. In *2011 International Conference on Electrical and Control Engineering* (2011), IEEE, pp. 5090–5093.

[177] Zhang, S., Tay, Y., Yao, L., and Liu, Q. Quaternion knowledge graph embedding. *arXiv preprint arXiv:1904.10281* (2019).

[178] Zhang, X., Chandrasegaran, S., and Ma, K.-L. Conceptscope: Organizing and visualizing knowledge in documents based on domain ontology. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021), pp. 1–13.

[179] Zhang, Y., Zhang, F., Yao, P., and Tang, J. Name disambiguation in aminer: Clustering, maintenance, and human in the loop. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2018), pp. 1002–1011.